Empirical studies in health and development economics

Deng, Zichen

2021

document version
Publisher's PDF, also known as Version of record

EMPIRICAL STUDIES IN HEALTH AND DEVELOPMENT ECONOMICS

ISBN: 978 90 3610 636 8

This book is no. 770 of the Tinbergen Institute Research Series, established through cooperation between Rozenberg Publishers and the Tinbergen Institute. A list of books which already appeared in the series can be found in the back.

VRIJE UNIVERSITEIT

EMPIRICAL STUDIES IN HEALTH AND DEVELOPMENT ECONOMICS

ACADEMISCH PROEFSCHRIFT

ter verkrijging van de graad Doctor aan
de Vrije Universiteit Amsterdam,
op gezag van de rector magnificus
prof.dr. V. Subramaniam,
in het openbaar te verdedigen
ten overstaan van de promotiecommissie
van de School of Business and Economics
op vrijdag 15 januari 2021 om 11.45 uur
in de online bijeenkomst van de universiteit,
De Boelelaan 1105

door

Zichen Deng

geboren te Jiangxi, China

promotor:	prof.dr. M. Lindeboom
copromotor:	dr. F.R.M. Portrait

CONTENTS

ACKNOWLEDGEMENTS

The journey started 15 years ago when I left my home town for the university. Wuhan, a very “famous” city this year, is around 300km from my hometown. Overwhelmed by the big city’s challenges, I managed to preserve the two most precious things from undergraduate life: economics and life partner. After spending several years in Shanghai, I started graduate study in Amsterdam. Over the years, I have met people all over the world. Their advice to my research (and life) is invaluable.

I want to start by thanking my supervisor Maarten Lindeboom for the supports in the last five years. Doing research is not easy and sometimes need great optimism with the challenges. Maarten was always optimistic and supported me in his special way. His guidance during the last five years has truly helped me to grow. I also need to thank France Portrait for being the copromotor.

The last chapter of the dissertation is a joint work with Sebastian. The project lasts very long, where we had ups and downs. But I really enjoyed the whole process. I also want to thank my other coauthors Jichun and Xianqiang, who share with me data sets since the start and valuable comments. Without these data sets and comments, I can’t finish all three papers by now. Of course, I am deeply grateful to all my coauthors, look forward to many more fruitful collaborations with them in the future. Yuanyuan Chen and Jimmy Chan were my supervisors back in Shanghai. Their helps in accessing data and short visits are extremely helpful.

Apart from my mentors and coauthors, the chapters of my thesis would not have been possible without the financial support from Vrije Universiteit and the EU Horizon 2020 research funding.

I want to thank Wendy Janssens, Erik Sørensen, Hans van Kippersluis, Gerard van den Berg, who, along with Jos van Ommeren, agreed to be members of my promotion committee and provided valuable comments.

One of the things that I like a lot during my Ph.D. life is attending conferences and workshops in different countries. I want to take the opportunity to thank all (both official and unofficial) discussants of my paper. Their suggestions on both presentations and paper themselves are invaluable to me. I also want to thank all editors and referees who handled my papers. Reading those reports (especially ones from referee 2) is one of the

most painful experiences in the past years. However, needless to say, I also learn a lot from those reports.

I want to thank my colleagues at VU for all the lunches, events, and discussions. I want to mention specifically: Bas, Jose, Pieter, Iman, and Paul for all kinds of supports: internal seminars, mock interviews, comments on the paper, and suggestions on the presentations. I want to thank Nadine for writing a letter. Tips about applications and interviews are also invaluable.

I also thank my peers at TI and VU for their help and friendship. An incomplete list includes Bo, Casper, David, Diana, Dieter, Hemon, Huaiping, Junze, Jurre, Lingwei, Martin, Rik, Shihao, Travels, Yan, Yeorim, Yuan, and Yue. I want to thank Elisabeth for going through many of my preliminary drafts, and Coen, for translating the summary into Dutch and introducing me to the Suriname food. Also, many thanks to Christina and Arianne and for being invaluable on the job market.

None of this would have been possible without the support from my family, especially my parents. Finally, Diandian and I met at the very beginning of this long journey. I dedicate this dissertation to her.

Zichen Deng
Bergen, Norway, October 2020

List of Figures

List of Tables

1

INTRODUCTION

This thesis contains two parts: Chapter 2 and 3 look into the long-run impacts of early-life conditions;[1] Chapter 4 evaluates the improvements in air quality and the enforcement of environmental regulation after introducing air pollution monitors in the city.[2] Two seemingly disparate topics, however, both look into the basic needs of human beings. The motivation is to make people healthy when they live well when they are educated, and they will be more productive (Przeworski, 1986). Hence, expenditures in clean air, food, and health represent an investment in the most valuable resource we have, the people themselves (Ohlin, 1938). This is the focus of many current public policies across the world. All three chapters in the thesis center around the implementation, the effectiveness, and the mechanism of policies which can provide those essentials. The thesis also benefits from the "credibility revolution" (Angrist and Pischke, 2010) that uses econometric tools to build causal relationships.

The next two chapters start from the famous fetal origins hypothesis, which was first proposed by Epidemiologist David Barker (Barker and Osmond, 1986; Barker, 1990). Extensive studies have documented that the period of gestation has significant impacts on the health, education, income for an individual ranging from infancy to adulthood. The hypothesis, although not initially proposed by economists, is quickly matched with the "credibility revolution" in the economic research (Angrist and Pischke, 2010). The fetal

[1] Chapters 2 and 3 are based on Deng and Lindeboom (2019, 2020).
[2] Chapter 4 is based on Axbard (2020).

hypothesis is tested extensively given that the in utero period is very well defined. Children subjected to an external shock in utero can often be compared to similar children born a little earlier or later who escaped the shock. However, several important pieces are missing from this vast literature, which is the focus of the Chapter 2 and 3 of the thesis.

The first challenge of the fetal literature is a lack of measure in individual exposure. Most research exploits exogenous variation in the early-life conditions and relates this to outcomes later in life to identify the causal effect credibly. The exogenous variation usually comes from a "natural experiment" that causes some individuals to be more likely to be affected by adverse conditions than others. Importantly, the "natural experiment" is abstracted to macro environment indicators, which can be the event's timing or the event's location. Researchers classify individuals to the "treatment" or "control" group by linking their birth information to the event's information. The long-run causal impact equals the difference in health later in life between these two groups. Some research moved beyond that by also using the intensity of the "natural experiment". The idea is to compare individuals from "high treatment" areas to "low treatment" areas, but this is often done using a parametric model. Therefore, a series of challenges emerge about the specification choice in econometric analyses. To the best of my knowledge, most studies choose linear functions due to simplicity, while there is very little justification for these choices. Ultimately, the research constraint is partly due to the lack of individual exposure measures to the "natural experiment".

Chapter 2 reexamines the long-run impact of early-life nutritional shortage caused by the Great Chinese Famine, the second most studied famine (after the dutch hunger winter). The Famine occurred from 1958 to 1961 and is considered "the worst famine in human history". There is a consensus in the literature that the Famine was a direct consequence of Mao's Great Leap Forward, an economic and social campaign led by the Communist Party from 1958 to 1961 (Kung and Lin, 2003; Meng, Qian, and Yared, 2015). However, existing studies often disagree with each other about the long-run impact on health later in life. Firstly, the heterogeneity in reduced-form findings might come from the differences in famine exposure (Van den Berg, Pinger, and Schoch, 2016). Secondly, the misspecification of the famine indicator in econometric analyses might be a second reason that conflicting results are often documented in the literature. To advance on this, Chapter 2 develops a new two-sample method and aims to obtain an estimate of the average causal effect of early-life hunger on health later in life.

We follow Van den Berg, Pinger, and Schoch (2016) and use individual hunger recall information to evaluate the strength of the association between the Famine indicator

and an actual hunger episode. The sample to estimate the strength are individuals aged above 13 during the Famine, which are older than the sample to look at long-run impacts. We find a nonlinear relationship between mortality rates, the commonly used famine indicator, and the individual experience. Therefore, we estimate a Two-Sample Instrumental Variable (TSIV) model with log-transformed mortality rates as the instrumental variable. To decrease the model dependency in the two-sample setting, we preprocess the data and homogenizes different samples using a nearest-neighbor matching algorithm. Estimates from the model show that early-life hunger leads to an increase in the risk of diabetes, hypertension, and obesity for females, while no effects are found for males.

Since the start of the literature (Almond, 2006; Van den Berg, Lindeboom, and Portrait, 2006), much of early "fetal origins" work has focused on demonstrating the impact of extreme, traumatic experiences (disease outbreak, recessions, famines, severe environmental shocks, etc.) in early life. The evidence from different countries is strikingly consistent. More and more recent research interests have shifted to understand the long-run impacts of exposure to a purposeful large-scale distribution of resources ("mild" shocks, Almond, Currie, and Duque, 2018). Recent researches add new evidence that a wide range of factors in the early years matter to child development. However, slightly different from the early literature, which looks that extreme experiences, there is often considerable heterogeneity in the effects of specific "mild" shocks. Some of the heterogeneity may be due to parental responses that either exacerbate or mitigate the effects. Understanding of the heterogeneity is, therefore, the research theme of the Chapter 3 of the thesis.

Chapter 3 investigates the long-run impacts of a nationally implemented salt iodization program on school-aged children's cognition in China. The national program follows the worldwide campaign initialed by WHO to eradicate Iodine Deficiency Disorders (IDD). The policy was very ambitious that all the salt in the market is regulated to contain iodine since late 1994.[3] We digitized biennial goiter data sets collected by local health agencies and the World Health Organization (WHO). The goiter data is then linked to recent survey data sets. Exploiting the policy's timing and historical iodine deficiency variations across provinces, we look into the causal impact of iodine deficiency in utero. The analysis focuses on in-utero exposure as the medical literature relates iodine deciency in the first trimester to impaired neurodevelopment. We find strong positive program effects on human capital formation for girls. Yet, we do not see any improvement in cognition for boys. There are reasons to believe the biological difference can

[3] 22 (out of 1800) counties were exempted due to high iodine content in the drinking water.

1

explain our findings. However, such a "nuclear" explanation is not consistent with all the existing literature.

Figure 1.1: High-end Iodized Salt Sold in China (2018)

Notes: Made in Friesland using seawater from the Wadden Sea (Waddenzee).

Our empirical strategy, similar to others used in the literature, interacts with the sudden introduction of the new salt (blip 1) with a short critical period (blip 2).[4] The compact nature of these two blips is what helps to estimate the causal effect. The kind that the treatment is well defined is always good to have a credible identification. The simplicity of the empirical design is so attractive that we, as economists, often forget that human beings respond to incentives, shocks, and public policies. Parents can reinforce or compensate for shocks by increasing or decreasing their investments. The investments can be monetary investments but, maybe importantly, can also be the time investments to their children.

An appealing explanation of the gender difference in our empirical findings is the difference in parental investments. The differences are deeply rooted in gender norms, which are important in East and South-East Asia and the Middle East and North Africa. A vast literature has shown that females from these areas receive fewer investments from parents and are likely not to reach their full potential in education, health, and personal autonomy. The second part of the Chapter 3 is the first attempt to link the literature on

[4] Bleakley (2010b) terms the strategy "a blip-blip" design.

gender preferences with the research on the long-term impacts of early-life conditions. We show that gender differences begin early in life. Meanwhile, our evidences add evidence on how fetal disadvantage is unfolded over the life cycle.

A simple model of the human capital formation with parental investments helps rationale the gender differences in program effects. Parents optimally choose their investments based on their preferences, budget constraints, and the human capital production function. The reduced-form program effect is broken down into the biological effect and the behavioral effect. The behavioral effect includes parental efforts that might mitigate or reinforce the biological effect. Our empirical finding is in line with the behavioral effect hypothesis that may differ by gender. Before the salt iodization policy, parents may have countered initial adverse shocks for boys and less so for girls. Therefore, when boy preferences are relevant, girls may benefit more from nationally implemented programs. Salt-iodization program might crowds out private parental investments in cognition.[5]

The simple model of human capital formation highlights three fundamental elements of the literature aiming at boosting human capital formation: parental preferences, budget constraints, and the production of human capital. Past interests have been mainly on identifying the production function, which measures the effectiveness of the parental input. Some recent studies use either RCT or conditional cash transfer to exploit the variation in parental resources, essentially the variation in budget constraints. However, parental preferences were largely ignored by the literature. Chapter 3 aims to fill this gap. A direct conclusion from the study is that large scale programs can have positive (and possibly) unintended effects on gender equality in societies with son preference. More importantly, any policy design on human capital formation should carefully consider behavioral responses from parents.

Like food and nutrition discussed in the first two chapters, clean air is another necessity for human beings. However, new pollution levels in emerging economies like China and India exceed the highest levels ever recorded in rich countries. Heavy air pollution reduces lifespans, decreases productivity, and has long-term impacts of early childhood exposure to air pollution on adult outcomes. Despite ambitious environmental laws in many countries worldwide, the enforcement of these regulations is often weak. Holding government officials accountable for this lack of enforcement is, in turn, often marred by inadequate information about environmental quality.

Chapter 4 studies whether better environmental monitoring can solve this issue and improve the effectiveness of the policy. The paper uses the introduction of a nation-

[5] Other consequence of the policy is that parents might divert their investments into other skill dimensions, notably for boys. Indeed, we find some program effects on non-cognitive skills for boys, but not for girls.

1

Figure 1.2: Countries with National Environmental Laws

Year	Countries with national environmental framework laws
1972	Norway, Sweden, United States
1992	Algeria, Armenia, Azerbaijan, Belarus, Bolivia, Brazil, Bulgaria, Canada, China, Colombia, Congo, Czechoslovakia, Democratic People's Republic of Korea, France, Gambia, Germany, Greece, Guatemala, Guinea, India, Indonesia, Iran, Iraq, Ireland, Italy, Jamaica, Kazakhstan, Kyrgyzstan, Latvia, Libya, Lithuania, Luxembourg, Madagascar, Malaysia, Mali, Malta, Marshall Islands, Mauritius, Mexico, Micronesia, Netherlands, New Zealand, Nigeria, Norway, Oman, Pakistan, Palau, Papua New Guinea, Paraguay, Peru, Philippines, Poland, Portugal, Republic of Korea, Russia, Saint Kitts and Nevis, Samoa, Senegal, South Africa, Sri Lanka, Swaziland, Sweden, Switzerland, Tanzania, Thailand, Togo, Tunisia, Turkey, Turkmenistan, Ukraine, United Kingdom, United States, Uzbekistan, Venezuela, Yugoslavia, Zambia
2017	Afghanistan, Albania, Algeria, Angola, Antigua and Barbuda, Argentina, Armenia, Australia, Austria, Azerbaijan, Bahrain, Bangladesh, Belarus, Belize, Benin, Bhutan, Bolivia, Bosnia and Herzegovina, Brazil, Brunei Darussalam, Bulgaria, Burkina Faso, Burundi, Cabo Verde, Cambodia, Cameroon, Canada, Central African Republic, Chad, Chile, China, Colombia, Comoros, Congo, Costa Rica, Côte d'Ivoire, Croatia, Cuba, Czech Republic, Democratic People's Republic of Korea, Democratic Republic of the Congo, Denmark, Djibouti, Dominican Republic, Ecuador, Egypt, El Salvador, Equatorial Guinea, Eritrea, Estonia, Eswatini, Ethiopia, Fiji, Finland, France, Gabon, Gambia, Georgia, Germany, Ghana, Greece, Guatemala, Guinea, Guinea-Bissau, Guyana, Haiti, Honduras, Hungary, Iceland, India, Indonesia, Iran, Iraq, Ireland, Italy, Jamaica, Japan, Jordan, Kazakhstan, Kenya, Kiribati, Kuwait, Kyrgyzstan, Laos, Latvia, Lebanon, Lesotho, Liberia, Libya, Liechtenstein, Lithuania, Luxembourg, Madagascar, Malawi, Malaysia, Maldives, Mali, Malta, Marshall Islands, Mauritania, Mauritius, Mexico, Micronesia, Mongolia, Montenegro, Morocco, Mozambique, Myanmar, Namibia, Nepal, Netherlands, New Zealand, Nicaragua, Niger, Nigeria, Norway, Oman, Pakistan, Palau, Panama, Papua New Guinea, Paraguay, Peru, Philippines, Poland, Portugal, Qatar, Republic of Korea, Republic of Moldova, Romania, Russia, Rwanda, Saint Kitts and Nevis, Samoa, San Marino, Sao Tome and Principe, Saudi Arabia, Senegal, Serbia, Seychelles, Sierra Leone, Singapore, Slovakia, Slovenia, Solomon Islands, South Africa, Sri Lanka, Sudan, Sweden, Switzerland, Syria, Tajikistan, Tanzania, Thailand, The former Yugoslav Republic of Macedonia, Timor-Leste, Togo, Tonga, Trinidad and Tobago, Tunisia, Turkey, Turkmenistan, Tuvalu, Uganda, Ukraine, United Arab Emirates, United Kingdom, United States, Uruguay, Uzbekistan, Vanuatu, Venezuela, Viet Nam, Yemen, Zambia, Zimbabwe

Notes: This graph shows countries with national environmental laws (UN, 2019). Data comes from the Environmental Law Institute.

wide pollution monitoring program in China to investigate the impact of installing additional monitors on both pollution concentration and government enforcement of pollution regulations. 177 cities were required to set up air quality monitors in 2015. The size of each city determines the number of air pollution monitors. Satellite measured air pollution is lower for cities with more monitors after monitors were set up in 2015. Since most regulations are set at the national level, all these cities should comply with the same governmental law. It's reasonable to believe that the difference in program effects is the policy implementation difference, which depends on local officials.

Empirical analysis of new enforcement data precisely confirms this conjecture. The reduction in air pollution is arguably driven by a significant increase in regulatory enforcement activities. Exploiting georeferenced enforcement records, we find that firms' growth is driven by firms located within 20km from a monitor, whereas firms beyond 20km face no change in enforcement. The geographic pattern echos the UN report last year that lack of enforcement is one of the most significant obstacles to combat climate change and pollution.

An examination of possible mechanisms shows that local officials' career concerns can explain a big part of our findings. Mayors who are about to retire put minimal efforts in regulating industrial firms. Consequently, effects on the pollution reduction of better monitoring disappear when the mayor of the city is retiring. We also document

that when monitors are insulated from government influence and external parties are responsible for providing information, the effect of an additional monitor on enforcement and reductions of pollution are both more substantial.

Both Chapter 3 and 4 show that policies with good purposes can generate unintended (not necessarily harmful) impacts. Universal Salt Iodization (USI), which targets the total population, decreases gender inequality as parents compensate more for boys. Similar to the adjustment of parental investment to their children, politicians respond to the introduction of new information. The mayors with incentives shift their focus from the economic development to the environmental target. And firms which happened to reside close to those monitors now are bearing additional abatement costs. Our data indeed shows the decrease in air pollution with extra monitors. Still, we also need to keep in mind that such a policy imposes different firms' standards from separate locations. The discretion in enforcement practice distorts resource allocation, which might harm the long-run economic development. All these findings can be traced to the 4^{th} principle of "ten principles in economics" from Econ101: people respond to incentives.

2

HUNGER [1]

2.1. INTRODUCTION

A large number of observational studies have demonstrated that health and economic disparities may have roots early in life. This relationship has been shown studies that examine the association between birth weight and later-life health and socioeconomic outcomes, as well as from studies that use "natural experiments" that cause some individuals (the treated) to be more likely to be affected by adverse conditions than others (the controls). In natural experiments, researchers mostly use contextual factors at an aggregate level to proxy for individual circumstances early in life. Examples of natural experiments include epidemics (Almond, 2006), economic conditions (Van den Berg, Lindeboom, and Portrait, 2006) and famines (Chen and Zhou, 2007; Lumey, Stein, and Susser, 2011). This paper adds to this literature by looking at the long-run consequences of actual exposure to hunger early in life.

In empirical applications, there are often limits to using aggregate indicators as a proxy for individual conditions. First, being born when the event in question took place is not equivalent to actually being exposed to adverse conditions. For example, researchers have used famine in a region as a proxy for being exposed to hunger. However, living in a food-deprived area is not equivalent to actually experiencing hunger, even if the famine's timing and location are precisely known. Wealthier households may still have sufficient food, or some parts of an exposed area may be less affected by the famine. Also, there is

[1] This chapter is based on Deng and Lindeboom (2020).

often uncertainty about the location and timing of the famine. For instance, the famine may be preceded by a prolonged period of food insecurity, so it is not always clear when the famine started. Most studies rely on historical evidence about the famine's evolution, and in some cases (such as the Chinese Famine), there may be conflicting information from historical sources. In these cases, researchers have limited information to justify the choices underlying their empirical approach, how they define the "treatment" period, which famine indicator they use, or the functional form of the parametric model. Without information on actual exposure among the survivors, the estimates are, at best, attenuated intention-to-treat (ITT) effects, but they may be biased.[2]

One solution is to use more granular data about hunger prevalence, such as excess mortality rates or food prices. But these may also be imperfect proxies for the nutritional environment, since they may not include informal trade systems (for prices) or other contextual factors that affect mortality rates and exaggerate (or attenuate) the intensity of the famine. The other alternative is to use actual hunger experience. Recently, this information has become available in a few data sets, such as the European Household Survey on Retirement and Aging (SHARE) and the Chinese Family Panel Survey (CFPS). These data resolve uncertainty about the precision of famine proxies and provide causal evidence about the treatment effect of actual hunger exposure on later life outcomes. The data richness is a significant advantage when studying the long-run impacts of early childhood conditions. In this paper, we use hunger recall information to estimate the effect of undernourishment early in life on health, measured by the Metabolic Syndrome Index (Hoynes, Schanzenbach, and Almond, 2016). We develop a Two-Sample Instrumental Variable (TSIV) method that relaxes the homogeneity assumption in standard TSIV methods and can deal with two samples from different populations.

In a recent paper, Van den Berg, Pinger, and Schoch (2016) developed a TSIV approach to examine the causal effect of early-life hunger exposure on later life health outcomes among SHARE respondents from 3 European countries. They estimate the strength of the association between the famine and actual hunger, and the effect of hunger on height. Hunger recall information is also imperfect because those experiencing hunger at very early ages have much less hunger recall than respondents who were older during the famine. To avoid this problem, Van den Berg, Pinger, and Schoch (2016) use another sample of older siblings with more reliable hunger recall information to measure the association between the famine and hunger experience.

The idea of combining two samples can be traced back to Angrist and Krueger (1992)

[2] For example, a weak correlation between the aggregate indicator and actual exposure, which is often caused by model misspecification, will bias the parameter estimates.

and was further developed in Inoue and Solon (2010). Two-sample methods implicitly assume homogeneity between the two samples.[3] However, this homogeneity assumption may not always be satisfied in practice. Very young children are, in general, frailer than older children, adolescents, and prime-aged adults. Therefore, very young children surviving a famine are more likely to have better biological traits or to come from more affluent families than their older counterparts. As a consequence, for the primary sample of those who are exposed early in life, the distribution of observed and unobserved confounding factors,is likely to be different from the distribution in the second (auxiliary) sample of older individuals. If there is heterogeneity in confounding factors, our estimates of the causal effect will be biased when the parametric model is misspecified.

Van den Berg, Pinger, and Schoch (2016) solve this by using discrete instruments and covariates and stratifying the samples into a finite set of homogeneous subsamples. When the famine's start and end are not precise, or if there is substantial variation in the famine's intensity across regions, one may want to rely on continuous instruments such as excess mortality rates or prices. Researchers would prefer a method that accommodates continuous instruments and covariates.

To combine information from two different samples in a robust way, we propose combining non-parametric preprocessing during the first step with a second-step regression estimator, as in Ho, Imai, King, and Stuart (2007). The first step decreases the model dependency on functional forms of the parametric causal inference in the second step. We derive properties of the proposed post-matching sample estimators and show that the estimator in the matched sample asymptotically recovers the actual regression coefficient in the primary (target) sample. Importantly, our results are valid for continuous treatment variables and continuous instruments. Monte Carlo simulations show that, when preprocessing restores homogeneity across the two samples, classical two-sample methods return unbiased estimates, while in the raw (i.e., not preprocessed) data, both the first stage and the second stage effects are biased.

Using the two-step method, we reexamine the long-run health impact of the Great Chinese Famine. The Great Chinese Famine has been studied extensively. Early studies (Chen and Zhou, 2007; Almond, Edlund, Li, and Zhang, 2010) documented substantial effects on height, wealth, and cognitive function in later life. On the other hand, there is conflicting evidence from more recent studies.[4] These differences may be the result

[3] Recent developments (Zhao, Wang, Spiller, Bowden, Small et al., 2019) of the two-sample model have pointed out the importance of this implicit homogeneity. They focus on heterogeneity between distributions of the instrumental variables in two samples. In practice, the distributions are usually similar enough in both samples when instruments are defined at an aggregate level (as in our study).

[4] For example, the study by Kim, Fleisher, and Sun (2017) uses nationally representative data and finds that famine exposure has no effects on later-life chronic conditions, while studies based on regional representative

of different studies exploiting different historical sources and using (slightly) different instruments. We focus on individuals born during or shortly before the famine (1957–1962).[5] Unlike previous Chinese famine studies, we use hunger recall information and supplement the primary data set with a second sample of much older individuals born between 1910 and 1947. For these older cohorts, hunger recall error biases are less of a problem. We use this second (auxiliary) sample to estimate the effect of famine exposure on the probability of reporting hunger.

We use nearest-neighbor matching to homogenize the distributions of covariates in the two samples and examine the relationship between famine exposure and the probability of reporting hunger. Virtually all previous papers implicitly assume a linear relationship between famine indicators and actual hunger exposure among the survivors.[6] We find that the linear approximation for the first stage does not fit the data and requires a logarithmic transformation of excess mortality rates (EDR) to make the relationship linear. This nonlinear relationship between EDR and hunger experiences has consequences for the previous contributions in the Chines Famine literature that estimated reduced-form models that were linear in the instruments. Next, we estimate the impact of hunger on an index of metabolic syndrome and find that early-life hunger for females leads to a 0.4 standard deviation increase in later life metabolic syndrome. This number is much larger than previous estimates in the literature. For males, we find much smaller and insignificant effects.

Our analysis makes four important contributions to the literature on long-run effects. We are the first study on the Chinese famine to provide evidence on the strength of the famine-hunger association. Our first stage results also validate the commonly used instruments in this literature to proxy undernourishment. We find a strong association between the aggregate famine indicators and reported hunger experience. However, this association is not a simple linear relationship. This nonlinearity might explain some of the conflicting findings in the literature.[7] Second, we make a methodological contribution to the two-sample instrumental variable model. We show that heterogeneity in confounded covariates in the auxiliary and primary samples can bias estimates of the

data (see, e.g., Huang, Li, Wang, and Martorell (2010), Li, Jaddoe, Qi, He, Lai, Wang, Zhang, Hu, Ding, Yang et al. (2011)) find significant results. Meng and Qian (2009) and Xu, Li, Zhang, and Liu (2016) also find null results.

[5] We use individuals born after the famine, 1963–1967, in a placebo test in section 2.5.4.

[6] Depending on the data set and the empirical design, the famine is usually approximated by province-level mortality rate or projected county-level cohort loss after the famine. These instruments in the literature are included linearly in the reduced-form regression equation.

[7] Studies looking at provinces with high mortality rates during the famine might find a very small effect, since the mortality rate is not informative for hunger at high mortality rates. We show a weak first-stage (F-stat < 10) when using a linear function in section 5.2.

causal effect. We propose using a non-parametric method to process the data before the parametric econometric analyses. Our simulation results show that estimates are less model-dependent when preprocessing balances the primary and the auxiliary samples.

Third, we provide new evidence on the long-run impact of early-life hunger experiences. All but one of the previous papers in this area used reduced-form approaches that include famine indicators rather than actual hunger experience. They thus estimate intention-to-treat (ITT) effects. We show that in the Chines famine, the causal treatment effects are much larger than the intention-to-treat effects found in this literature. Finally, our paper discusses the exclusion restriction required in IV famine studies. In our context, where we aim to assess the effect of undernourishment and use hunger recall information, we have to assume that the famine affects children only via hunger, and that there are no other channels. Although the famine's primary impact is food restriction, we cannot exclude other potential impacts such as stress and/or infectious diseases that often accompany famines.[8] Thus, it is likely that the exclusion restriction is violated. Violations of the exclusion restriction are relevant when interpreting reduced-form ITT estimates. We adopt a recently proposed exercise that bounds the treatment effect under weaker assumptions (i.e., that relaxes the strict exclusion restriction Conley, Hansen, and Rossi, 2012). Our bounding exercises shed light on the nutrition contribution of the famine's impact relative to all other potential channels. We conclude that our main results hold under much weaker assumptions.

The remainder of the paper is structured as follows. Section 2.2 briefly describes the historical background, institutional setting, and important features of the Great Chinese Famine. In section 2.3, we describe the data sets we use in this study and discuss our malnutrition indicator and our outcome variables. Section 2.4 introduces our empirical model, discusses our identification assumptions, and addresses potential threats to identification. Section 2.5 presents our main results.

2.2. BACKGROUND AND PRIOR RESEARCH

2.2.1. THE GREAT CHINESE FAMINE

The Great Chinese Famine occurred from 1958 to 1961 and is widely considered "the worst famine in human history". During the Famine, at least 16.5 million individuals perished in rural areas.[9]

[8] The famine lasted a few years, and during a prolonged period of undernourishment, the disease environment may change.

[9] See Sen (1981) and Ravallion (1997) for estimates of total famine casualties. More detailed evidence can be found in Yang (1998); Yao (1999); Meng, Qian, and Yared (2015).

Since 1949, the central government adopted the Stalinist development model, which emphasized investment in industrial sectors. The rural sector had to provide resources for investment and raw materials for production. To accommodate high investment in industry, the government initiated a large scale land reform, followed by an aggressive collectivization policy. During the land reform period (1950–52), redistribution of landlord-held land and other property boosted agricultural production. Major indicators of productivity in the rural sector, such as grain and cotton outputs, had double-digit growth rates during this period. Collectivization of the rural sector followed immediately after this rapid growth period. It started with the "Five-Year Plan" (1953–57), in which peasant households were organized into agricultural producers' cooperatives. This reform dramatically slowed agricultural growth rates.

On the eve of the famine, the central government in China controlled food production, distribution, and consumption. Approximately 80% of the population worked in the agriculture sector. Grain was harvested and stored communally, and private stores of grain were prohibited. The central government procured grain produced in rural areas from communal depots after the fall harvest. Procured grain was fed to urban workers, exported to other countries in exchange for industrial equipment and expertise, and stored in reserves as insurance against natural disasters. The grain retained by the rural regions was used to feed the peasants in communal kitchens, which were established so that the collective could control the preparation and consumption of food. Furthermore, the government prevented peasants from migrating and, consequently, peasants could only consume the food distributed to their collective.

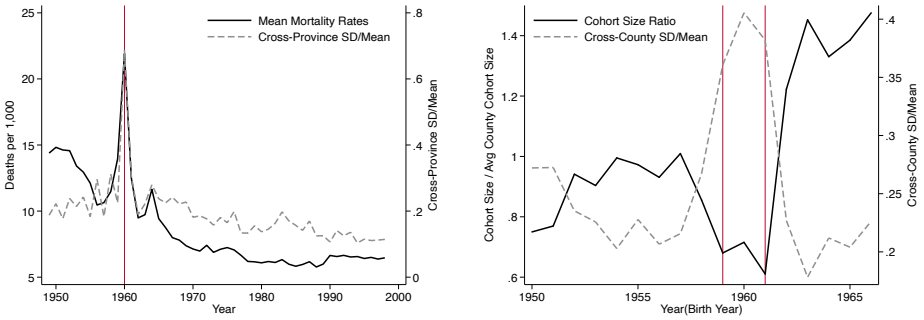
There is a consensus in the literature that the Great Chinese Famine was a direct consequence of Mao's Great Leap Forward, an economic and social campaign led by the Chinese Communist Party from 1958 to 1961 (Kung and Lin, 2003; Meng, Qian, and Yared, 2015). During the campaign, the political climate encouraged provincial leaders to overstate grain production and even export grain to signal the success of Mao's Great Leap Forward (see Meng, Qian, and Yared, 2015). Despite a severe shortage of food, China was a net grain exporter in 1960 (Yao, 1999; Lin and Yang, 2000).

2.2.2. RELEVANT FEATURES OF THE FAMINE

The famine lasted until 1962, but some researchers have argued (see Tan, Zhibo, and Zhang, 2015, for an example) that birth and death rates in some provinces had already returned to normal levels by 1961. The precise end date for the famine is not clear for all provinces. With hunger recall data from the CFPS, we can address this issue in more detail (see section 2.3). The famine also featured considerable variation in severity across

regions. In 1960 death rates for two adjacent provinces could differ by more than five-fold. For instance, in 1960 the province of Anhui had a death rate of 1.84%, while the neighboring province of Jiangsu had a death rate of 0.29%.

Figure 2.1: Average and Spatial Variation in Famine Severity



(a) Province-level mortality rates (b) County-level survivor birth cohort sizes

Notes: Figure 2.1a: The solid line plots mean mortality rates, which are average mortality rates across provinces in each year. The dashed line is the standardized variance in mortality rates across provinces in year t. Figure 2.1b: The solid line plots the de-trended 1% size of the birth cohort born in year t. The dashed line is the normalized cross-county variance in birth cohort sizes. Source: Meng, Qian, and Yared (2015)

To better depict the variation in severity across regions during the Great Chinese Famine, we present some graphical evidence from Meng, Qian, and Yared (2015). Figure 2.1a plots average mortality rates and the normalized variance in mortality rates over time (the cross-province standard deviation divided by the cross-province mean). The figure shows that during the famine (denoted by the two vertical lines), both mean mortality and the variance in mortality rates spiked. Our empirical strategy exploits the variation in famine induced mortality rates across provinces. Figure 2.1b provides complementary county-level evidence. The figure plots mean and cross-county standardized variance in cohort size.[10] This figure shows a clear drop in cohort size and increased variance during the famine.

2.2.3. SELECTED FAMINE STUDIES

For an overview of famine studies, we refer to Van den Berg and Lindeboom (2018). Here we highlight the results from the two most widely studied famines: the "Dutch Hunger Winter" and the Chinese Famine. The Dutch Hunger Winter (December 1944–April 1945) is the most studied famine in the epidemiological and demographic literature on long-

[10] Note that the cohort size reflects mortality as well as the fertility effects of the famine. See also 2.5.4.

run effects. The Dutch Hunger Winter has a number of features that are advantageous for researchers: it arrived unexpectedly, lasted for a short period, and took place in a relatively stable society with thorough data collection. See Lumey, Stein, and Susser (2011) for an excellent review of studies that used this famine. Studies using this famine found effects on blood glucose levels, diabetes, severe obesity, high blood pressure (hypertension), and schizophrenia. (Ravallion, 1997; Roseboom, de Rooij, and Painter, 2006; Stein, Kahn, Rundle, Zybert, van der Pal-de Bruin, and Lumey, 2007; Scholte, Van den Berg, and Lindeboom, 2015) find negative effects on labor market outcomes and hospitalization outcomes for those exposed in the first trimester of gestation.

The Great Chinese Famine is perhaps the second most used famine in the literature on the long-run effect of exposure early in life. Li and Lumey (2017) provides an extensive review and meta-analysis of the medical and epidemiological research. They conclude that the literature has found effects for overweight, type 2 diabetes, hyperglycemia, metabolic syndrome, and schizophrenia. However, most studies vary substantially in exposure definition, control selection, and analytical methods. Consequently, when controlling for these differences, they conclude that "most effects commonly attributed to the famine can be explained by uncontrolled age differences between exposed and control groups".[11] One of the earliest economics papers (Chen and Zhou, 2007) finds substantive effects for height, labor supply, and earnings. Their findings are confirmed by Meng and Qian (2009), who addressed measurement error of famine exposure by exploiting unique institutional determinants. (Almond, Edlund, Li, and Zhang, 2010) uses data from the China Population Census 2000 to look at the effect of famine exposure on literacy, labor market status, wealth, and marriage market outcomes. They find that exposed women marry later and have less educated spouses. They also find evidence for the Trivers-Willard hypothesis that the sex-ratio of the offspring of exposed parents favors daughters. Few economic studies target specific chronic conditions such as diabetes and hypertension. One exception is Kim, Fleisher, and Sun (2017), who do not find effects on chronic diseases such as hypertension.

2.3. DATA

2.3.1. HUNGER RECALL

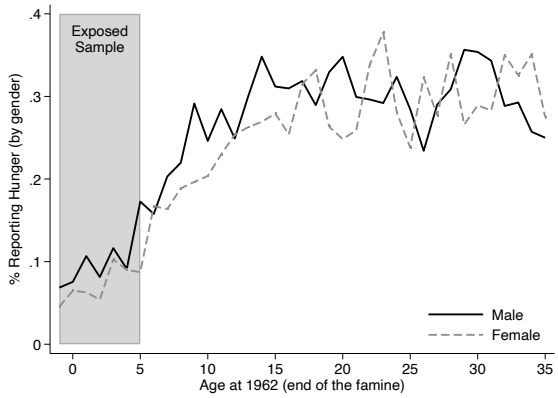
Our main data come from the China Family Panel Study (CFPS), a large-scale, nationally representative panel survey conducted by the Institute of Social Science Survey at Peking University. Currently, four waves are available, 2010, 2012, 2014, and 2016. The

[11] This is in line with Xu, Li, Zhang, and Liu (2016), who find that estimates of the famine effects are sensitive to the choice of health indicators, measures of famine severity, and regression model specifications.

baseline wave (hereafter CFPS-2010) is collected through a multistage probability sampling procedure and consists of 14,798 households. All adults living in the household are interviewed, leading to a total sample of 34,425 adult observations.

Similar to the SHARE survey, the CFPS-2010 survey included a question on hunger recall. The survey asked: "Have you experienced starvation for more than one week? If so, when did it start, when did it end, and where did it happen?" Since the question only requires the experience to last more than one week, we don't know how many weeks in total respondents have experienced food shortage in each year.[12] A critical aspect of this hunger experience measurement is that the question did not explicitly mention the Great Chinese Famine. The question only asked about general hunger experiences. The non-response rate for the hunger experience question is very low at 0.055.[13] Most of the hunger experiences happened during the Great Chinese Famine, which occurred more than 50 years before the survey. Hunger responses related to the great Chinese famine are likely to be subject to recall bias, especially for respondents born close to the famine. Figure 2.2 supports this suspicion.

Figure 2.2: Probability of Reporting Hunger Conditional on Famine Exposure



Notes: The exposed sample includes individuals who were born between 1958–1962.

The figure displays the fraction of individuals in the raw data who report hunger as a fraction of those who were alive during the famine period. The horizontal axis is the respondent's age in 1962; the vertical axis is the fraction of individuals who experienced hunger during the famine (1958–1962). The fraction of hunger recall increases with age

[12] Ideally, we would consider a "dose-response" specification if a more refined measure of hunger were available. Instead of a dummy treatment variable, we also use the number of years exposed to hunger before age five. The findings are similar.

[13] This is close to the non-response rate in the SHARE survey.

during the famine and stabilizes at about 30 percent after age 12.[14] We see no gender differences in reports of hunger during the famine period. Our primary sample contains individuals born before and during the famine, whose own recall of malnutrition around birth and in the first years of life is likely to suffer from significant recall bias. To overcome this problem, we follow the idea introduced by Van den Berg, Pinger, and Schoch (2016) to use recall information from individuals who experienced the famine at an older age to proxy for actual hunger exposure for individuals in the primary sample.

2.3.2. THE PRIMARY AND AUXILIARY SAMPLE AND SUMMARY STATISTICS

In the primary analysis, we restrict the sample to include individuals born between 1958 and 1962 and who lived in rural areas. After dropping individuals with missing information on the outcome variable (discussed below), we are left with 958 males and 972 females from 27 provinces. Very few people migrated during the famine (Chen and Zhou, 2007). This is also true for our sample of people living in rural China: only 4% live outside their province of birth at the time of the interview. For the reported health outcome, we pool data sets from the first three waves of CFPS (CFPS-2010, CFPS-2012, and CFPS-2014).[15] We look at three chronic diseases: hypertension, diabetes, and obesity. The first two conditions are derived from a question also used in other social surveys: "Has a doctor ever told you that you suffer from…". Obesity is defined as a body mass index (BMI) exceeding 29 ($BMI > 29$). We used self-reported height and weight to calculate BMI. We construct a Metabolic Syndrome Index by grouping information on all three chronic conditions. As discussed by Kling, Liebman, and Katz (2007); Hoynes, Schanzenbach, and Almond (2016), aggregating multiple measures to an index improves statistical power. The index is the average of the standardized z-scores for each component.[16] High values of the index are associated with worse health.

To construct an auxiliary sample that is less susceptible to recall bias, we select females born between 1910 and 1947, aged 15–52 in 1962. If the female in the auxiliary sample is the mother of the individual in the primary sample, we include this observation as a proxy for the individual's hunger exposure. We found that 10% of matches are between the individual from the primary sample and his/her mother. For the remaining

[14] Several validation studies match recall data with actual outcomes and find that the recall data is reliable when individuals reach adult ages (see, for example, O'malley, Bachman, and Johnston (1983) on teen drinking behavior).

[15] The fourth wave, CFPS-2016, does not collect information on some of the health conditions we need for our health index (hypertension, diabetes, and height.

[16] We calculate the z-score by subtracting the mean and dividing by the standard deviation. We calculate both the mean and standard deviation using the primary analysis sample (i.e., individuals born between 1957 and 1962 in all three waves of CFPS).

individuals in the primary sample, we match on the village of birth (or if not available, county of birth, or province of birth) and next on age and literacy.

Table 2.1: Summary Statistics

	Female sample			Male sample		
	Obs.	Mean	SD	Obs.	Mean	SD
Panel A: Primary sample - basic information						
Age at 2010	958	50.40	1.88	972	50.59	1.88
Mother literate	958	0.15	0.36	972	0.18	0.39
Mother birth year	958	1931.39	7.26	972	1931.03	7.67
Panel B: Primary sample - health outcomes						
Hypertension	2522	0.04	0.19	2618	0.02	0.15
Diabetes	2522	0.01	0.10	2618	0.01	0.09
Obesity	2522	0.04	0.21	2618	0.04	0.19
Metabolic syndrome(index)	2522	0.02	0.64	2618	-0.02	0.54
Panel C: Auxiliary sample						
Mother literate	3682	0.24	0.43	3682	0.24	0.43
Mother birth year	3682	1925.88	15.58	3682	1925.88	15.58

Notes: Author's tabulations of CFPS-2010, CFPS-2012, and CFPS-2014. Panel A summarizes background information for individuals born in rural area between 1957 and 1962. Panel B pools chronic conditions data from three waves of CFPS. The Metabolic Syndrome Index is the z-score from subtracting the mean and dividing by the standard deviation. Both the mean and standard deviation are calculated using the analysis sample (individuals born between 1957 and 1962 in all three waves of CFPS). High values of the index are associated with worse health. Panel C displays the background information in the auxiliary sample, which includes all individuals born prior to 1947.

We report summary statistics for the primary and auxiliary samples in Table 2.1. Panel A reports summary statistics in the primary sample for the main outcome and age and literacy status for the individual's mother. We report summary statistics for the health outcomes collected in CFPS2010, CFPS-2012, and CPFS-2014 in Panel B. Panel C reports summary statistics for the auxiliary sample (born between 1910 and 1947). We use the same auxiliary sample for the male and female sub-samples. Comparing the primary sample with the auxiliary sample, we see that mothers' literacy rates in the primary sample are much lower than mothers' literacy rates in the auxiliary sample. We also see that mothers' age in the auxiliary sample is about six years older than the mothers of individuals in the primary sample. This invalidates classical two-sample IV methods. Below we extend our two-sample IV method so that we can apply it when the two samples have different distributions of observed and unobserved characteristics.

2.4. TWO-SAMPLE IV MODELS WITH HETEROGENEOUS SAMPLES

Many researchers have used data combination methods to identify the causal effect when no single sample contains all relevant variables (see Ridder and Moffitt, 2007, for a review). Most empirical applications of two-sample methods implicitly assume homogeneity between the primary sample and the auxiliary sample. Table 2.1 in section 2.3 showed substantial differences in age and literacy status between the (proxy) mothers in the primary and auxiliary samples. Ignoring that these samples differ in important ways will result in first-stage estimates that are not relevant for the primary sample and thus irrelevant for the treatment effect in the second stage (in the primary sample). We consider a framework that is similar to Van den Berg, Pinger, and Schoch (2016):

$$Y_i = \psi(D_i, X_i, U_i), \tag{2.1}$$

where D_i denotes severe hunger during childhood for individual i. Y denotes health in adulthood. X denotes a vector of observed covariates. We are interested in the causal effect of hunger experiences (D) in early-life on later-life outcomes (Y). There are a number of challenges to identifying the causal effect: D is likely to be endogenous, and D is systematically misreported or not in the same data set as Y. The most widely used approach is to find a contextual factor (Z) that is an instrument for D and estimate the intention-to-treat (ITT) effect. For instance, researchers have used being born in a famine-stricken area or excess mortality rates in an area as an instrumental variable for undernourishment early in life.

In practice, we set up a parametric framework, e.g. a linear model, to approximate the true causal model (2.1):

$$Y_i = \gamma D_i + \pi X_i + U_i, \tag{2.2}$$

where i indexes the individual. γ is the causal effect of hunger early in life on later-life health, our parameter of interest. When the model has one endogenous variable and one instrumental variable, the estimates of γ consist of two components: the reduced-form (or ITT) estimates (2.3)

$$Y_i = \gamma_0 Z_i + \pi_0 X_i + W_i; \tag{2.3}$$

and, when hunger experience information (D_i) is available, the first stage regression (2.4)

$$D_i = \gamma_1 Z_i + \pi_1 X_i + V_i. \tag{2.4}$$

Most papers in the famine literature use the linear function (2.3) to estimate ITT effects. Note, however, that this linearity assumption implicitly assumes a linear relationship between the treatment variable D_i and the instrumental variable Z_i. This assumption may not be innocuous. When hunger information D is available, we can use two-stage least squares (TSLS) estimators. From Vansteelandt and Didelez (2018) and Buja, Brown, Berk, George, Pitkin, Traskin, Zhang, and Zhao (2019), we know that the TSLS estimator is consistent for γ even if the linear instrument-exposure model is misspecified. Even when the relationship between the treatment and the instrument is non-linear, we can still recover γ by separately estimating two linear equations (2.3) and (2.4).

The TSLS estimate of the local average treatment effect is equivalent to the ratio of the reduced-form and first-stage estimates: $\gamma = \gamma_0/\gamma_1$. With variables in two different samples, we can estimate two equations separately: we use the primary sample to estimate the reduced-form equation and the auxiliary sample to estimate the first-stage equation. In our study, we use the sample of children born during the famine to estimate the reduced-form equation, and the auxiliary sample of adults during the famine to estimate the first-stage equation. Although TSLS is robust to model misspecification in the one-sample setting, this robustness property does not carry over to the two-sample setting (TSIV, Angrist and Krueger, 1992). A recent literature (Zhao, Wang, Spiller, Bowden, Small et al., 2019; Shu and Tan, 2020) investigates a similar question, the impact of heterogeneity between distributions of the instrumental variable in two different samples. Usually this is not a problem in economic studies that use natural experiments, since the instrument is taken at an aggregate level.[17] Here we are concerned with heterogeneity in the distribution of covariates (observed and possibly unobserved) between the two samples.

Van den Berg, Pinger, and Schoch (2016) solve the problem of heterogeneous samples by stratifying the data into homogeneous subsamples and then integrating over a finite set of possibilities to obtain a non-parametric Wald estimator of the treatment effect. This approach is convenient in their application, where both the treatment variable and the instrument are binary. In situations like ours, where information on the start and end of the famine is unclear and there is variation in intensity across regions, researchers may want to rely on continuous instruments Z_i such as excess mortality rates (EDR) or prices. Continuous instruments make it impossible to apply their non-parametric estimator in these situations.

[17] For instance, in the literature on long-run effects of early childhood conditions, contextual factors at the regional level are used to instrument for individual conditions early in life. The problem of heterogeneity in the distribution of instruments is likely to be more relevant in genomics applications, where a specific genomics instrument is at the individual level.

In our study, we use children born during or shortly before the famine to estimate equation (2.3), and people who were already adults to estimate equation (2.4). Mortality profiles are generally U-shaped, implying higher mortality rates for the very young and the very old, and low mortality rates for adults in their prime years. Therefore, the mortality impact of a famine at very young ages is likely to be different from the mortality impact at adult ages. The famine will, therefore, differently affect cohorts and thus result in differential mortality selection across different cohorts. Further, young children surviving the famine are more likely to come from families with favorable biological traits and/or from (wealthier) families who had better access to food. Table 2.1 showed that there were substantial differences between the distribution of covariates in the primary sample and the auxiliary sample. Two-sample estimates using the original (i.e., raw unbalanced) samples are therefore biased. In Appendix A, we illustrate the bias of TSIV estimates using simulation exercises.

Two-sample instrumental variable (TSIV, Angrist and Krueger, 1992) and two-sample two-stage least squares (TSTSLS, Inoue and Solon, 2010) are two widely used parametric estimators. Following Ho, Imai, King, and Stuart (2007), we propose a two-step approach to address the issue of heterogeneous samples. In the first step, we employ nonparametric preprocessing, such as nearest-neighbor matching, to balance the covariate distributions between the primary sample and the auxiliary sample. In the second step, we perform a parametric analysis using the primary sample and the matched individuals from the auxiliary sample.

The non-parametric preprocessing in the first step decreases the dependence on parametric modeling assumptions in the second step. The simplest way to understand our approach is to consider one-to-one-exact matching. This matches each individual in the primary sample to a close match in the auxiliary sample. After the matching procedure, the preprocessed auxiliary sample is balanced with the primary sample, with any unmatched auxiliary units discarded. With all units in the primary sample matched, this procedure eliminates dependence on the functional form of the parametric analyses in the second step. As a result, misspecification in the second step is less likely to be a source of bias. This type of two-step procedure is often called “doubly robust”.

To formalize this idea, we show that the matched post-matching two-sample estimator is consistent. In the first step, two samples are matched based on observables X.[18] With the matched sample, a two-sample IV model is estimated following equations (2.3) and (2.4). Let R_i be an indicator variable equal to 1 if the ith unit is drawn from the

[18] In practice, the matching can be based on different sets of observables X in each step. However, to simplify notation, we use the same observables X in both steps.

primary population, and 0 otherwise. Now suppose that we are able to observe D in the target population (primary sample), and let $\beta = [\gamma_1, \pi_1]$ be the vector of regression coefficients from regression (2.4) in the target population (i.e., the primary sample). Similarly, let $\tilde{\beta} = [\tilde{\gamma_1}, \tilde{\pi_1}]$ be the vector of regression coefficients obtained from regressing D on Z and X in the matched sample.

Assumption 1 (Random Sampling)**.** *The primary sample, $\mathscr{S} = (Y_i, Z_i, X_i)_{i=1}^{N_0}$, is a sample obtained from N_0 draws from the population distribution of (Y, D, Z, X). The auxiliary sample, $\mathscr{S} = (D_i, Z_i, X_i)_{i=1}^{N_1}$, is a sample obtained from N_1 draws from the population distribution of (Y, D, Z, X).*

Like Abadie and Imbens (2012), we assume that the data consist of a random primary sample and a random auxiliary sample to simplify the discussion. Let S^* be the matched sample generated by matching each unit in the primary sample, i, to the auxiliary sample, $J(i)$ without replacement. We only consider one-to-one matching, since the auxiliary sample in our empirical application (section 2.5) is only slightly larger than the primary sample. We choose the sets of matches $J(i)$ to minimize the sum of the matching discrepancies, $\sum_{i=1}^{N_0} d(X_i, X_{j(i)})$. Similar to the matching literature, we assume that the sum of matching discrepancies vanishes quickly enough to allow asymptotic unbiasedness as the sample size (N) increases.

Assumption 2 also ensures that the support of the common variables (X) in the primary population is contained within the support in the auxiliary population.

Assumption 2 (Support Condition)**.** *Let $\mathscr{X}_1 = supp(X|R=1)$ and $\mathscr{X}_0 = supp(X|R=0)$, then $\mathscr{X}_1 \subseteq \mathscr{X}_0$.*

We now describe the population distribution targeted by the matched sample, S^*. Let $P(\cdot|R=1)$ and $P(\cdot|R=0)$ be the source distributions of (Y,S) that we draw the primary and auxiliary samples in S from, and let $E[\cdot|R=1]$ and $E[\cdot|R=0]$ be the corresponding expectation operators. For given $P(\cdot|R=1)$ and $P(\cdot|R=0)$ and a given number of matches M, we dene a matching target distribution, P^* , over the triple (Y,S,W), as

$$P^*((Y,S) \in A|R=1) = P((Y,S) \in A|R=1) \text{ and}$$
$$P^*((Y,S) \in A|R=0) = E[P((Y,S) \in A|R=0,X)|R=1].$$

The first expression holds because the primary sample is our matched targeting distribution. The second expression, the distribution of (Y,S) in the auxiliary sample, is generated by integrating the conditional distribution of (Y,S) given X and $R=0$ over the distribution of X given $R=1$ in the primary sample.

Assumption 3 (Propensity Score Equality). $P(D=1|Z,X,R=0) = P(D=1|Z,X,R=1)$

Assumption 3 requires predictive invariance for the treatment between the two heterogeneous populations. This assumption is similar to the idea of selection-on-observables.

Proposition 1. *Under regularity conditions, the regression coefficients ($\tilde{\beta}$) of D on Z and X in the matched sample, S^*, asymptotically recover the analogous regression coefficients (β) in the population of the primary sample.*

Proposition 1 formalizes the idea that the first-stage estimates of the matched sample recover the parameters of the matching target distribution (i.e., the distribution of the primary sample). We prove Proposition 1 in Appendix A. Note that our results are valid for continuous treatment variables (D) and continuous instruments (Z).

We illustrate this proposition using Monte Carlo simulation.[19] The covariates are drawn from uniform distributions in both the primary and auxiliary samples. The support of the distribution is designed to be larger in the auxiliary sample. Therefore, the simulated covariates are heterogeneous across the two samples. Additionally, the relationship between the endogenous variable and the instrumental variable depends on covariates. Without knowing the actual data generating process, estimates from a heuristic (misspecified) linear model using the auxiliary sample are an average effect over the population in the auxiliary sample, which is likely to be different from the average effect over the population in the primary sample. The simulation results show that both the first-stage and second-stage estimates are biased using the original (unprocessed) data. The non-parametric matching procedure restores homogeneity across samples, and after this preprocessing, classical two-sample IV returns unbiased estimates.

2.5. REEXAMINING THE LONG-RUN EFFECT OF CHINESE FAMINE

In the previous section, we developed a two-step procedure to combine information from heterogeneous samples. In this section, we apply the method to estimate the causal effect of famine induced undernourishment early in life on later-life health. The Great Chinese famine has been studied extensively, and all studies use regressions like (2.3). These studies resulted in different and sometimes conflicting conclusions. This discrepancy is largely due to differences in instruments Z, different data sets, different selections made in the construction of the analysis sample, and different specifications for the reduced-form regression (2.3), see Li and Lumey (2017).

[19] We discuss the details of the simulation in Appendix A.

2.5.1. NEAREST-NEIGHBOR MATCHING

We use nearest-neighbor matching based on village of birth, mother's age, and literacy to balance the two samples. When village of birth information is not available, we match individuals on county of birth, ensuring that the matched pairs have similar family background characteristics. Cao, Xu, and Zhang (2020) has documented that counties with large family clans experienced lower mortality during the famine. Table 2.2 presents the balanced primary and auxiliary samples after applying nearest-neighbor matching. As an out-of-sample test, we also show summary statistics for two father's characteristics, which the matching algorithm does not target. The percentage of literate fathers is balanced between the primary and auxiliary samples. Our matching algorithm significantly improves balance for the average age of the father. The mean age difference decreases from 5 to 1.5 after the matching procedure. This improvement signals that the balance between the primary and auxiliary samples has been improved significantly, even for variables we did not explicitly target.

Table 2.2: Summary Statistics – Matched Sample

	Female sample			Male sample		
	Obs.	Mean	SD	Obs.	Mean	SD
Panel A: Matched primary sample						
Age at 2010	956	50.40	1.88	970	50.59	1.88
Mother literate	956	0.15	0.36	970	0.18	0.39
Mother birth year	956	1931.37	7.26	970	1931.04	7.68
Panel B: Matched auxiliary sample						
Mother literate	956	0.13	0.33	970	0.16	0.37
Mother birth year	956	1930.63	10.12	970	1930.43	10.35

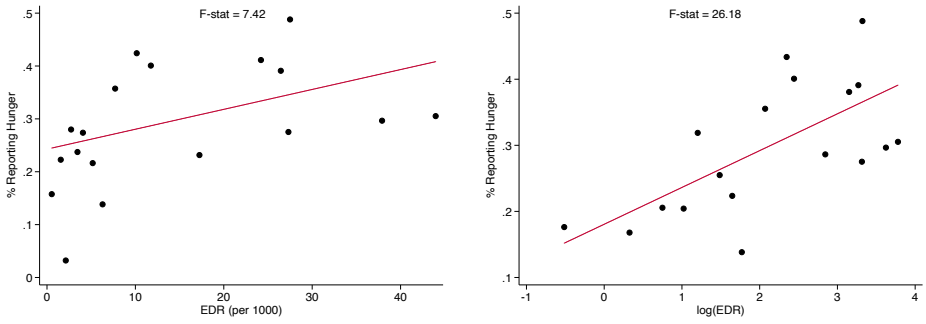
Notes: Author's tabulations of CFPS-2010. Panel A summarizes background information for individuals in the matched primary sample. Panel B summarizes background information for individuals in the matched auxiliary sample.

2.5.2. RESULTS FROM TWO-SAMPLE IV MODELS

The excess mortality rate (EDR) is commonly used in studies of the Chinese Famine (Chen and Zhou, 2007; Almond, Edlund, Li, and Zhang, 2010), as well as in other studies (for instance, Bleakley (2007) use it as a measure of disease prevalence). We take province-level mortality rates from Meng, Qian, and Yared (2015) and construct our instrumental variable (i.e., the excess death rate in 1960) following Chen and Zhou (2007). We define the excess death rate in 1960 as the gap between the death rate in 1960 and

the average death rate in the three years before 1959.

Figure 2.3: Mean Hunger Experiences versus EDR/log(EDR)



(a) Mean Hunger Experiences versus EDR (b) Mean Hunger Experiences versus log(EDR)

Notes: Figure 2.3a and 2.3b present two binned scatter plots of the relationship between the percentage of respondents who had hunger experiences during the famine and excess mortality rates in 1960. The excess death rate in 1960 is the gap between the death rate in 1960 and the average death rate in the three years before 1959. Figure 2.3a uses raw excess mortality rates, while Figure 2.3b uses the log-transformation of excess mortality rates. The points on the figure plot the mean hunger experiences within each EDR/log(EDR) percentile bin. The best-fit line is estimated using an OLS regression on the underlying micro data. F-statistics for both regressions are reported separately.

We first check that our instrument is relevant for the excess mortality rate. Figure 2.3a presents a binned scatter plot of the relationship between hunger experiences and excess mortality rates. The points on the figure plot the percentage of respondents who had hunger experiences during the famine. Interestingly, a linear relation, which is implicitly assumed by most famine studies, does not do a good job capturing the relationship between EDR and hunger. The percentage of individuals who reported any hunger experience remains more or less stable at 40% for excess mortality rates above about 15. This implies that, for high excess mortality rates, the instrument is no longer informative for actual hunger exposure. Indeed, the associated F-test of the linear regression is 7.42, suggesting that the (linear) EDR is a weak instrument. This may explain why studies that restrict their analysis to provinces with high mortality rates generally find small effects.[20]

An arbitrary solution would be to restrict the sample by leaving out provinces with extremely high mortality rates. However, this will considerably decrease the sample size and may generate selection bias.[21] Instead, we decided to take a log-transformation of excess mortality rates in the province where the individual lived (log(EDR)). Figure 2.3b

[20] A large proportion (7/10) of regional studies surveyed in Li and Lumey (2017), which look at high mortality areas, finds insignificant impacts.

[21] The marginal survivor at high mortality rates will be different from the marginal survivor at lower mortality rates.

shows that the relation between hunger experiences and the log(EDR) can be captured well by a linear function. The associated F-statistic for the regression is well above 10.

Table 2.3: First-stage – the Effect of Famine Intensity on Hunger

	(1)	(2)	(3)	(4)	(5)	(6)
	Hunger					
	All	All	Female	Female	Male	Male
log(EDR)	0.075***	0.075***	0.087***	0.087***	0.062***	0.062***
	(0.015)	(0.015)	(0.015)	(0.015)	(0.016)	(0.016)
Mother literate		-0.020		0.0025		-0.044*
		(0.026)		(0.041)		(0.023)
Age		-0.00094		-0.00081		-0.0022
		(0.0064)		(0.0090)		(0.0099)
Observations	1926	1926	956	956	970	970
F-Stat	26.18	26.18	32.92	32.92	14.93	14.93

Notes: Each parameter is from a separate regression of hunger between 1957–1962 on log(EDR) (EDR is short for excess death rates). We estimate the model on the matched auxiliary sample. The stratification by gender is based on the gender variable in the primary sample. Standard errors, clustered by province of the birth, are in parentheses. *, **, *** indicates significance at the 10%, 5% and 1% level, respectively.

Table 2.3 presents our estimation results from the first-stage linear regressions based on the auxiliary sample, with an indicator for hunger (recall) as the dependent variable. The table includes regressions without controls (columns 1, 3, and 5) and regressions with controls (2, 4, and 6). As controls, we include age and literacy status for the (proxy) mother. The table also presents estimates for the full sample (cf. Figure 2.3b) and estimates by gender. Across all specifications, we find highly significant effects of the instrument (log(EDR)) and F-statistics that well exceed 10, indicating that there is no problem of weak instruments (Staiger and James, 1997).

Table 2.4 presents TSIV estimates of the treatment effect on the matched primary sample. The table shows that famine-induced hunger increases the standardized metabolic syndrome index by about 0.38. The ITT estimate from the reduced-form regression (2.3) equals 0.032 (see Table B2 of Appendix B for the full table). For males, the coefficients are small and insignificantly different from zero. These findings are at odds with a number of studies on the long-run health effects of adverse conditions in early childhood. Most of these studies find boys to be more sensitive to adverse conditions early in life. Our results are, however, in line with findings from other studies of the Great Chinese Famine. See e.g. Almond, Edlund, Li, and Zhang (2010) who also find stronger effects for females. Table B1 in Appendix B shows results for the separate components of the metabolic syn-

Table 2.4: Effects of Hunger at Age 0–5

	(1)	(2)	(3)	(4)
	Metabolic syndrome (index)			
	Female	Female	Male	Male
Hunger before age 5	0.37***	0.38***	0.045	0.062
	(0.14)	(0.14)	(0.20)	(0.20)
Mother literate		0.028		0.031
		(0.034)		(0.026)
Age		0.013***		0.0050
		(0.0041)		(0.0043)
Observations	2517	2517	2612	2612

Notes: The results are based on TSIV estimates from separate regressions. All regressions use the log(EDR) as the instrumental variable on the (matched) primary sample of individuals born between 1957 and 1962 from three waves of the China Family Panel Survey (CFPS). Standard errors clustered by province appear in parentheses. *, **, *** indicates significance at the 10%, 5% and 1% level, respectively.

drome index. For females, we find significant effects for all individual components (albeit only at the 10% level). Interestingly, for males we only find a significant effect on obesity.

A significant proportion of previous studies that examine the Chinese Famine use the linear EDR in regressions like (2.3) as an instrument/proxy for the severity of the famine. To highlight that our two-step two-sample method's is robust to functional form misspecification, we also present the instrumental variable estimate using the linear EDR in Table B3 of Appendix B. The table shows that all four IV estimates are very close to our main estimates using the log-transformed EDR (Table 2.4). Due to the weak instruments, the standard errors are much larger. However, the similarity in the IV estimates' magnitudes in Table B3 and Table 2.4 provides additional support for our preprocessing. Specifically, these results support our claim that homogeneity between the two samples decreases biases due to model misspecifications.

2.5.3. ROBUSTNESS TO VIOLATIONS OF PERFECT EXOGENEITY

One limitation of our instrumental strategy is that the exclusion restriction required for a causal interpretation of the IV estimates may not be satisfied in our context.[22] Famines may be accompanied by increased stress. Epidemiological studies find that prenatal stress exposure in humans is associated with later-life health outcomes, in partic-

[22] This is less of a problem in famine studies that estimate reduced-form regressions that include famine indicators to obtain ITT effects on later life health. However, for these studies, the exclusion restriction assumption is relevant if one wants to interpret the ITT effect as being driven solely by undernourishment.

ular memory problems, decreased learning, depression, and dementia (Selten, van der Graaf, van Duursen, de Wied, and Kahn, 1999; Heffelfinger and Newcomer, 2001). Further, during a prolonged period of undernourishment the disease environment may change, which may increase the prevalence of infectious diseases. This may affect later-life health and socioeconomic outcomes (Almond, 2006). So, while we expect hunger to be the most important channel through which a famine affects later life outcomes, it is likely that alternative channels directly affect later-life health as well.

To examine the robustness of our TSIV estimates, we relax the assumption of perfect exogeneity and derive bounds on the true effect of malnutrition early in life on later-life health following Conley, Hansen, and Rossi (2012). We only perform this analysis for females.[23] Consider a generalization of the standard IV model that allows the instrument Z to enter linearly in the second-stage regression,

$$Y_i = \gamma D_i + \lambda Z_i + \pi X_i + U_i. \tag{2.5}$$

Conley, Hansen, and Rossi (2012) shows how to obtain consistent estimates of the effect of interest (here γ, the effect of famine-induced undernourishment on health) if λ is known. By choosing a set of fixed values for $\lambda = \lambda_0$ and running separate regressions for each value of $\lambda = \lambda_0$, we can evaluate the sensitivity of the IV estimate of γ to violations of the exclusion restriction.[24] Conley, Hansen, and Rossi (2012) choose the ITT estimate of the reduced-form regression as the relevant values of $\lambda = \lambda_0$. In our case these values come from equation (2.3). With an ITT estimate of 0.032 (see Table B2 of Appendix B) for females, we examine the sensitivity of our TSIV estimate for $\lambda \in [0, 0.02]$.

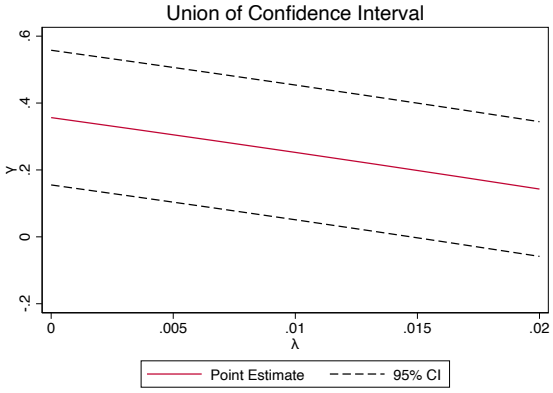
We plot the TSIV estimates for different values of λ in Figure 2.4. The (straight) red line is the point estimate of γ under different values of λ. The dotted line is the 95% confidence interval. $\lambda = 0$ corresponds to the assumption that the famine affects later life health only via hunger (D). Indeed, the point estimate at $\lambda = 0$ is about 0.38. Higher values of λ are associated with more important alternative channels and hence smaller coefficients for hunger. For instance, with λ equal to 0.016, we assume that only 50% of the famine affects health via undernourishment.

Figure 2.4 shows that the straight line of point estimates for γ is relatively flat. Specifically, at $\lambda = 0.016$, γ is about 0.2 and still significantly different from zero. The IV estimate only becomes insignificantly different from zero for higher values of λ. To reject the long-run impact of undernourishment, one must assume that at least 50% of the effect of the

[23] The previous section only showed significant effects for females.

[24] Another approach to finding this sensitivity is to use a sub-sample for which the instrumental variable does not affect the treatment variable (van Kippersluis and Rietveld, 2018). In our setting, urban cohorts in the famine are one candidate. However, since the urban cohort is so small, we don't pursue this approach.

Figure 2.4: 95% Confidence Intervals to Exclusion Restriction Violations



Notes: The graph shows how the IV estimate of the effect of hunger in early life on hypertension changes when the exclusion restriction $\lambda = 0$ is violated. Estimates and confidence intervals come from estimating Equation (2.5) for females at different xed values $\lambda = \lambda_0$. The red line is the point estimate under different values of lambda; the dotted line is the confidence interval. We control for fixed effects for year of birth, province of birth and year of interview. Family controls include mother's literacy and age.

famine is due to channels other than undernourishment. Overall, this exercise indicates that our results are robust to moderate violations of the exogeneity assumption. The strength of the first-stage estimates may be important for this finding. Documented in Conley, Hansen, and Rossi (2012), minor deviations from the exclusion restriction may greatly decrease precision when instruments are weak. Table 2.3 showed that the first stage estimates were strong and robust when additional covariates are included. Figure 2.4 shows that the benefit of having a strong instrument is that biases will be small when perfect exogeneity is violated.

2.5.4. ADDITIONAL ROBUSTNESS CHECKS AND SELECTION ISSUES

In this section, we explore some additional specification checks to examine the robustness of our findings. Table 2.5 presents these additional specifications. Columns 2 and 4 present estimates of hunger's effect with additional controls (age and literacy status of the mother) in the regression; columns 1 and 3 report estimates without additional controls.

Our mortality rate information is at the province level, so we use standard errors clustered at the province of birth in our main specification. We also consider clustering at the village of birth. We present these results in Panel A of Table 2.5. Clustered standard errors by the province of birth are in parentheses; standard errors clustered at the village

Table 2.5: Robustness

	(1)	(2)	(3)	(4)
	Metabolic syndrome (index)			
	Female	Female	Male	Male
Panel A: alternative clustering				
Hunger before age 5	0.37***	0.38***	0.045	0.062
	(0.14)	(0.14)	(0.20)	(0.20)
	[0.20]	[0.20]	[0.28]	[0.29]
Observations	2517	2517	2612	2612
Panel B: small sample window				
Hunger before age 5	0.44***	0.45***	0.059	0.074
	(0.12)	(0.12)	(0.22)	(0.22)
Observations	1859	1859	2040	2040
Panel C: control for migration				
Hunger before age 5	0.36***	0.36***	-0.014	-0.0070
	(0.13)	(0.13)	(0.20)	(0.20)
Observations	2387	2387	2534	2534
Panel D: placebo test using cohort 1964-1967				
Hunger before age 5	-0.22	-0.23	-0.26	-0.21
	(0.20)	(0.20)	(0.35)	(0.36)
Observations	2707	2707	2519	2519

Notes: Each coefficient is from a separate regression. All regressions use the log(EDR) as the instrumental variable. In columns (2) and (4), we control for mother's literacy and age. In Panel B, we drop individuals born in 1962. In Panel C, we drop individuals for whom the residing province is different from the province of birth. In Panel D, we estimate the same model using individuals born between 1964 and 1967. Standard errors clustered by province appear in square brackets. *, **, *** indicates significance at the 10%, 5% and 1% level respectively.

level are in square brackets. Standard errors clustered at the village level are slightly larger than those at the province level, but the estimates for females remain significant at the 5% level. We also calculated standard errors clustered at the county of birth and found similar results.

We also use a slightly smaller sample by restricting our sample to individuals born between 1958 and 1961. Although recent studies (Tan, Zhibo, and Zhang, 2015) show that the food shortage problems in some provinces were still present in 1962, the earlier literature often assumed that the famine ended in 1961. An advantage of this age/cohort restriction may be that the instrument is less noisy. However, the smaller sample size may reduce statistical power. Panel B of the table shows that the effects for females be-

comes larger and the standard errors are actually somewhat smaller.

Third, we check the impact of migration. CFPS collects very detailed information on the province of birth. While across province migration is limited (only 4%), our matching algorithm may perform worse for individuals who migrated.[25] To check this, we drop 4% (82 cases) of individuals who do not live in their province of birth at the time of the survey. We report these results in Panel C of the table; our estimates are hardly affected by this restriction.

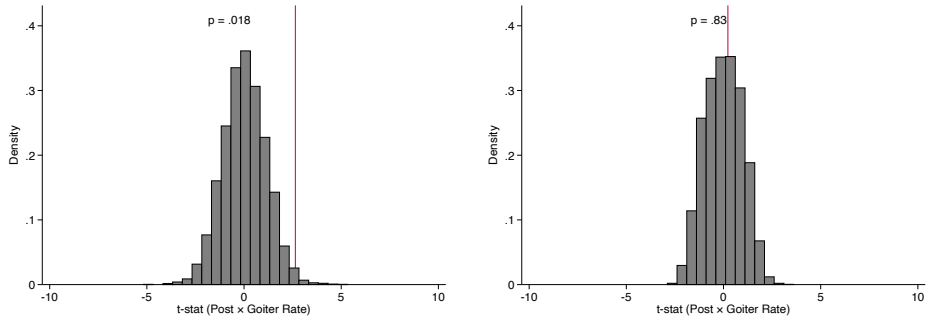
Additionally, we conduct a placebo exercise. Recall that we use variation in the peak excess mortality rate across provinces and link these mortality rates to individuals born in these provinces between 1958 and 1962. To examine whether other (province-level) confounding effects may drive our results, we apply our two-step estimator to the cohorts born between 1964 and 1967. The choice of the 1964–1967 cohort for the placebo exercise is motivated by the literature that uses a DiD design, where the cohort born after the famine is used as a control. We report the results of this exercise in Panel D. The estimates have the wrong sign, but they are all insignificant. These results support the idea that our findings are driven by famine-induced hunger and not other mortality-related confounding factors that vary by province.

Finally, we conduct falsification tests to demonstrate the statistical power of our inferences by assigning a pseudo-treatment. We randomly assign province of birth and thus log(EDR) to each respondent in the main sample (cohort 1958–1962). If our identification strategy is valid, estimates using these pseudo-samples should be centered around zero. In Figure 2.5, we plot the distribution of the t-statistics from 5,000 estimated pseudo-treatment effects for males and females. The two distributions are both centered around zero. To address our model's statistical power, we mark the location of the t-statistic of the baseline treatment effects in the distribution of pseudo-treatment effects in Table 2.4. We also report the share of the pseudo-treatment t-statistics that exceed the actual t-statistic of the baseline model (in absolute values). These p-values support our design and the statistical power of our exercises.

A famine, especially when it lasts for a prolonged period, leads to selective mortality and selective fertility. Starting with selective fertility, Roseboom (2010) show that during the Dutch Hunger Winter of 1944 about half of the women in exposed areas did not menstruate. For the same famine, Stein, Susser, Saenger, and Marolla (1972) found that fertility reductions primarily occurred in families from lower socioeconomic classes. Besides fertility leading to a selective sample of births, there might also be differential mortality. In utero, mortality is likely to occur more often among frail fetuses. Frailty may depend

[25] Non-migrants are more likely to be linked to a family residing in the same village.

Figure 2.5: Pseudo-treatment Effects



(a) Females (b) Males

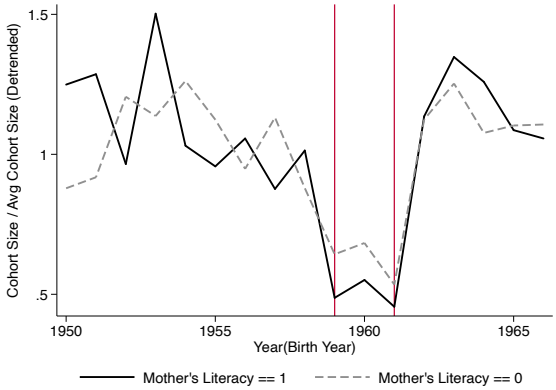
Notes: Pseudo-treatment vs. actual hunger exposure: the distribution of t-statistics resulting from 5,000 random assignments of treatment to individuals, as well as the t-statistics from the actual treatment through hunger exposure (red line). "p-values" report the share of the pseudo-treatment t-statistics that are larger than the actual t-statistics.

on biological traits or poor living conditions for women during the gestational period. In addition to in utero mortality, mortality may also occur between birth and the survey in 2010. The extent to which these selection effects take place will vary with the intensity of the famine. In the context of the Chinese famine, with substantial regional variation in famine intensity, this will lead to systematic differences in province populations, leading to biased inferences.

We examine the possible influence of the selection effects by looking at cohort sizes. We expand the CFPS data by including all individuals born between 1950 to 1966. We count the number of observations for each birth year and regress these cohort sizes on a linear time trend. Next, we use the ratio between observed cohort sizes and predicted sizes to plot the detrended cohort loss series. The resulting detrended series based on the CFPS closely resembles the detrended series based on the Census in 1990 (see Figure B1 of the Appendix B).

Figure 2.6 plots the detrended time series for literate and illiterate mothers. The figure shows a clear drop in cohort size of about 40–50 percent, indicating the importance of selective fertility and mortality. These cohort size reductions are in line with the findings in Roseboom (2010) and Scholte, Van den Berg, and Lindeboom (2015) for the Dutch Hunger Winter. It is important whether the cohort size-reduction differs by socioeconomic status. Figure 2.6 shows that there are only small differences in cohort size loss by socioeconomic status as measured by literacy status. This further supports the idea that our estimates are isolating the effect of hunger on later-life health.

Figure 2.6: Cohort Loss by Mother's Literacy



Notes: Figure 2.6 plots the detrended relative cohort sizes by mother's literacy using CFPS-2010.

2.6. CONCLUSION

Undernourishment early in life may have lasting effects on health and socioeconomic outcomes later in life. Most of the epidemiological and economic literature has produced intention-to-treat (ITT) effects from reduced-form regressions that relate later life outcomes to indicators of famine exposure early in life. This paper uses an indicator of *actual* hunger experience early in life from a hunger recall question in the China Family Panel Survey (CFPS). To estimate the effect of undernourishment on later life health, we develop a novel Two Sample IV (TSIV) approach that can deal with differences in the distribution of observable and unobservable factors in the two samples. Using the CFPS data, we find convincing evidence for long-term impacts on late-life chronic conditions of early-life malnutrition for females, but not for males.

Using hunger recall information has two clear advantages over previous studies. First, it allows us to examine the relationship between hunger experiences among survivors and the commonly used famine indicator (excessive death rates). This helps justify instruments used in this literature, as well as providing insight into the proper specification of the reduced-form regressions used in the extensive famine literature. Concerning reduced-form regressions, we find a non-linear relationship between hunger recall and the commonly used famine indicator (excess mortality rate). A linear specification leads to weak instrument problems. The previous literature on the Chinese famine, in essence, estimated linear relationships between the outcome of interest and famine exposure. This may, in part, explain the differences in findings across studies. Second,

with hunger recall information we can estimate the causal effect of undernourishment on later-life health. We show that these effects are robust to potentially mild and moderate violations of the exogeneity assumption.

The TSIV method developed in this paper can deal with differences in the distribution of observables and unobservables in the two samples. It first uses a non-parametric matching procedure to balance both samples, making the second-step less sensitive to misspecifications. This 'double robust' estimator is relevant in light of the recent interest in collecting and using recall information. Recall information, especially when it relates to experiences from ten or more years ago, may be subject to recall errors and may require auxiliary data to retrieve the relevant associations from the first stage equation in the TSIV model. In this framework, information on channels other than nutrition could be added and used to disentangle the relevance of different channels through which famines can affect later-life health. We leave this to future work.

A. ADDITIONAL THEORETICAL RESULTS

PROOF OF THE PROPOSITION 1

Proof. Let $Q(\cdot|R=1)$ and $Q(\cdot|R=0)$ represent the distributions of (D,S) given $R=1$ and $R=0$, respectively. We use the notation $D(0)$ to represent the latent D in the primary sample. Therefore, the regression coefficient in the primary sample, our target, is dened by $E_{Q(\cdot|R=1)}[(D(0)-K'b)^2]$.

$$E_{Q(\cdot|R=1)}[(D(0)-K'b)^2] = E_{Q(\cdot|R=1)}[E_{Q(\cdot|R=1)}[(D(0)-K'b)^2|X]] \tag{2.6}$$
$$= E_{Q(\cdot|R=1)}[E_{Q(\cdot|R=0)}[(D(0)-K'b)^2|X]] \tag{2.7}$$
$$= E[E[(D-K'b)^2|X,R=0]R=1] \tag{2.8}$$
$$= E^*[(D-K'b)^2|R=0] \tag{2.9}$$

In the last equation, the first equality follows from the law of iterated expectations; the second equality follows from propensity score equality. The last two equations follow the definition of the matching target distribution. Until here, we have shown that matching under propensity score equality allows us to reproduce the first stage setting for the primary sample. Therefore, the regression coefficient in the primary sample is recovered using the matched sample.

To further establish the large sample property of $\tilde{\beta}$ as $N_1, N_0 \to 0$, let $\tilde{\beta}$ be the vector of sample regression coefficients obtained from regressing D on $K=[Z,X]$ in the matched

sample,

$$\tilde{\beta} = \text{argmin} \frac{1}{N} \sum_{i \in S^*} (D - K'b)^2) = (\frac{1}{N} \sum_{i \in S^*} KK')^{-1} \frac{1}{N} \sum_{i \in S^*} KD. \tag{2.10}$$

From 6-9, the matching procedure makes sure $\frac{1}{N} \sum_{i \in S^*} KK' \xrightarrow{p} H$. $H = E(KK')$ is the Hessian, which is invertible by assumption.

$$\tilde{\beta} - \beta = (\frac{1}{N} \sum_{i \in S^*} KK')^{-1} \frac{1}{N} \sum_{i \in S^*} (KD - KK'\beta) \xrightarrow{p} 0 \tag{2.11}$$

□

To make sure that H is invertible, we need to a list of regulation conditions. Low-level assumptions can be found in Abadie and Imbens (2012).

To identify the treatment effect in a two-sample setting, we also need following standard assumptions in the instrumental variable model.

Assumption A1 (Instrument Relevance)**.** $Z \not\perp\!\!\!\perp D|X$

Assumption A2 (Instrument Independence)**.** $Z \perp\!\!\!\perp U|X$

Assumption A3 (Exclusion Restriction)**.** $Y \not\perp\!\!\!\perp Z|D,X,U$

With additional assumptions A1-A3, two-sample estimators (TSIV or TSTSLS) after matching can recover the local average treatment effect in the population of the primary sample.

SIMULATION STUDY

We investigate the finite-sample properties of our proposed two-step methods. The main focus is to evaluate the performance of the two-step approach we proposed. R codes for performing the simulations are provided in the Supplementary Material.

Let $\mathscr{U}(a,b)$ be the uniform distribution on $[a,b]$. R is an indicator variable, equal to 1 or 0 for the primary or auxiliary sample, respectively. To model the heterogeneity in two samples, $X|R=1 \sim \mathscr{U}(0,1)$ and $X|R=0 \sim \mathscr{U}(0,2)$.

For the primary population, we generate data according to

$$Y = 0.5D + X + U, \tag{2.12}$$

the endogenous variable D is defined as

$$D = 0.5XZ + X^2 + V, \tag{2.13}$$

where Z is distributed as $N(0,1)$ and (U,V) are distributed independently of (Z,X) as

$$\begin{pmatrix} U \\ V \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & 0.5 \\ 0.5 & 1 \end{pmatrix}\right). \tag{2.14}$$

For each simulation, we generated an i.i.d. sample of size $N_0 = 500$ and recorded only (Y,Z,X) from the primary population and generated an i.i.d. sample of size $N_0 = 1000$ and recorded only (Y,Z,X) from the auxiliary population. The two samples are then merged into one. We consider the following post-matching regression specifications, the first-stage equation:

$$D = \tau_1 Z + \beta_1 X + V; \tag{2.15}$$

the second-stage equation,

$$Y = \tau_0 D + \beta_0 X + U. \tag{2.16}$$

As we have one endogenous variable and one instrumental variable, we can estimate the model using either TSIV or TSTSLS. The specification is incorrect relative to the conditional expectation $E[D|Z,X]$ for the first-stage equation. For our simulation results, we will focus on estimators of τ_0. For the simulation, $\tau_0 = 0.5$ under the matching target distribution.

Table A1 summarizes the results based on 1,000 repeated simulations. In a regression that uses the full sample without matching, both TSIV and TSTSLS estimates of τ_0 are biased under misspecication. After matching, both TSIV and TSTSLS yield valid estimates for τ_0.

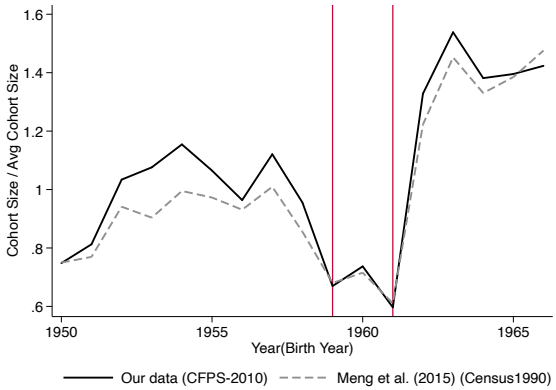
Table A1: Monte Carlo Results

	(1)	(2)	(3)	(4)
	TSIV		TSTSLS	
	Full sample	Post-matching	Full sample	Post-matching
τ_0	0.23	0.47	0.27	0.55
	(0.033)	(0.067)	(0.037)	(0.076)
τ_1	0.49	0.24	0.49	0.24
	(0.034)	(0.067)	(0.036)	(0.069)

Notes: Target parameter: coefficient $\tau_0 = 0.5$ and $\tau_1 = 0.25$. The results based on 1,000 repeated simulations. DGP can be found in Section A.

B. ADDITIONAL FIGURES AND TABLES

Figure B1: Cohort Loss in CFPS



Notes: The figure compares the relative survivor birth cohort sizes in our data set (CFPS-2010, the solid line) with the relative cohort sizes in Meng, Qian, and Yared (2015) (Census1990, the dashed line).

Table B1: Effects on Separate Components

	(1)	(2)	(3)	(4)
		Components		
	Metabolic syndrome (index)	Diabetes	High blood pressure	Obesity
Panel A: Female				
Hunger before age 5	0.37***	0.033*	0.075*	0.064*
	(0.14)	(0.020)	(0.039)	(0.035)
Observations	2517	2517	2517	2517
Panel B: Male				
Hunger before age 5	0.052	-0.034	-0.042	0.15***
	(0.20)	(0.034)	(0.064)	(0.049)
Observations	2612	2612	2612	2612

Notes: Each coefficient is from a separate regression. All regressions use the log(EDR) as the instrumental variable. The sample contains all individuals born between 1957 and 1962 in three waves of CFPS. Three components, diabetes, hypertension, and obesity, are dummy indicators constructed from CFPS. Standard errors clustered by province appear in square brackets. *, **, *** indicates significance at the 10%, 5% and 1% level respectively.

Table B2: Reduced-form Estimates at Age 0-5

	(1)	(2)	(3)	(4)
	Metabolic syndrome (index)			
	Female	Female	Male	Male
log(EDR)	0.031***	0.032***	0.0026	0.0035
	(0.011)	(0.011)	(0.012)	(0.012)
Mother literate		0.027		0.028
		(0.031)		(0.026)
Age		0.013***		0.0048
		(0.0044)		(0.0043)
Observations	2517	2517	2612	2612

Notes: The results are based on reduced-form estimates (2.3) from separate regressions. All regressions use the (matched) primary sample of individuals born between 1957 and 1962 from three waves of the China Family Panel Survey (CFPS). Standard errors clustered by province appear in parentheses. *, **, *** indicates significance at the 10%, 5% and 1% level, respectively.

Table B3: Effects of Hunger at Age 0-5

	(1)	(2)	(3)	(4)
	Metabolic syndrome (index)			
	Female	Female	Male	Male
Panel A: log(EDR) as the instrumental variable				
Hunger before age 5	0.37***	0.38***	0.045	0.062
	(0.14)	(0.14)	(0.20)	(0.20)
Mother literate		0.028		0.031
		(0.034)		(0.026)
Age		0.013***		0.0050
		(0.0041)		(0.0043)
Observations	2517	2517	2612	2612
Panel B: EDR as the instrumental variable				
Hunger before age 5	0.42*	0.43*	0.063	0.11
	(0.22)	(0.23)	(0.30)	(0.31)
Mother literate		0.021		0.031
		(0.033)		(0.028)
Age		0.013***		0.0052
		(0.0043)		(0.0045)
Observations	2517	2517	2612	2612

Notes: The results are based on TSIV estimates from separate regressions. All regressions are based on the (matched) primary sample of individuals born between 1957 and 1962 from three waves of the China Family Panel Survey (CFPS). Panel A uses the log(EDR) as the instrumental variable. Panel B uses the EDR as the instrumental variable. Standard errors clustered by province appear in parentheses. *, **, *** indicates significance at the 10%, 5% and 1% level, respectively.

3

IODINE [1]

3.1. INTRODUCTION

Iodine deficiency early in pregnancy can have significant, irreversible effects on the brain development of the fetus (Cao, Jiang, Dou, Rakeman, Zhang, O'donnell, Ma, Amette, DeLong, and DeLong, 1994) and can, therefore, have important consequences for human capital formation of children and subsequent socioeconomic outcomes. Indeed, the large medical literature (see Zimmermann, 2011, for a systematic literature review) has found adverse effects of iodine deficiency on later life growth and development, in particular when exposed early in life. Using historical data, recent economics papers such as Feyrer, Politi, and Weil (2017); Adhvaryu, Bednar, Molina, Nguyen, and Nyshadham (2020) found that cohorts exposed to higher levels of iodine intake have higher labor force participation rates, higher incomes and also higher probabilities of entering top-tier occupations with higher cognitive demands. Although the benefits of micro-nutrient improvement have been documented for decades, still only around 66% of households have access to iodized salt globally.[2] Further, little is known about pathways between early in life exposure and adult productivity and earnings. This impedes our understanding of how developmental trajectories unfold over the life course and who benefits most

[1] This chapter is based on Deng and Lindeboom (2019).

[2] In 2007, an estimated 31.5% of school-age children (266 million) had insufficient iodine intake. In the general population, this amounted to 2 billion people. See also: WHO, www.who.int/nutrition/publications/micronutrients/FNBvol29N3sep08.pdf, UNICEF, www.who.int/nutrition/topics/idd/en/.

from large scale interventions.

Our contribution to the extent of literature is to consider and account explicitly for possible interactions with gender preferences in receiving parental investments early in life. These differences are deeply rooted in gender norms, which are important in East and South-East Asia and the Middle East and North Africa. The idea that gender preferences play a role in the western world cannot be excluded either. Biased gender norms may already have an impact on children very early in life and might explain heterogeneous gender effects found in the literature on the long-run effects of early life shocks (Field, Robles, and Torero, 2009; Maccini and Yang, 2009; Adhvaryu, Bednar, Molina, Nguyen, and Nyshadham, 2020). This paper examines the effect of a massive, nationally implemented salt iodization program in China. Evidence of differential program effects by gender would not only provide new interpretation to gender differences in the impact of early in life shocks found in the literature but also shed light on predictions of theoretical models of human capital formation that allow parents to make compensatory and reinforcing investments in different dimensions of human capital. Parental child preferences play an important role in such investments. If compensatory investments of parents are relevant, then large scale interventions may crowd out (or reinforce) private investments and may also affect skills dimensions initially not aimed. Public policy can, therefore, have positive (and possibly) unintended effects on gender equality in societies where gender norms are important. Understanding mechanisms underlying program effects are crucial as any intervention would be blind without knowledge of the mechanisms underlying behavioral responses.

To fight against iodine deficiency-related diseases around the country, the Chinese government implemented a national program of regulating salt to contain iodine in October 1994. At the same time, biennial province-based monitoring was introduced to record the use and iodine content of household salt, along with urinary iodine concentrations among schoolchildren. After the introduction of the program, the urinary iodine concentration reached satisfactory levels from 1995 onward, and the percentage of children who had goiter[3] dropped rapidly. Given the importance of iodine during the gestational period for brain development (Cao, Jiang, Dou, Rakeman, Zhang, O'donnell, Ma, Amette, DeLong, and DeLong, 1994; Zimmermann, 2011), we, in the first instance, focus on the potential impact of this policy on cognitive ability and school attainment of children who are affected in utero.

[3] Iodine is an essential component of hormones produced by the thyroid gland. Iodine deficiency can lead to an enlarged thyroid gland located at the base of the neck, which is the most visible consequence of iodine deficiency. Goiter in adulthood does not have severe consequences, while iodine deficiency in utero can lead to impaired neurodevelopment of the child and post-birth cognitive disabilities.

For this, we link iodine deficiency information across locations collected at the start of the intervention to nationally representative rural samples drawn from the *China Family Panel Studies* (CFPS). A unique aspect of the CFPS survey is that math and vocabulary ability from standardized tests were collected besides information on educational attainment and schooling. Similar to Shah and Steinberg (2017), our human capital measurements in the CFPS have the advantage that the same questions were given to each individual in the survey, no matter whether he/she is currently enrolled in school or not. With this data set, our empirical design does not suffer from selection bias caused by censoring individuals who had already left school. To identify the long-term benefits of the salt iodizing policy, we use the national salt iodizing program as a quasi-experiment and exploit geographic variation in goiter prevalence before the intervention. So, we essentially compare improvements in math and vocabulary ability as well as educational attainment and years of schooling of cohorts conceived before and after the salt iodization in areas with varying pre-intervention goiter prevalence. Additionally, we make sure that we are comparing outcome variable trends (by birth cohort) across high and low goiter province of birth in their deviation from each region's average trend. We also include an extensive set of controls, such as pre-treatment province characteristics interacted with cohort dummies to flexibly control for confounding factors at the province level that might deferentially affect cohorts. Our difference-in-differences estimates show that the salt iodization policy has substantial and significant effects on cognition for girls. A one standard deviation (12%) decrease in the pre-intervention regional goiter rate is associated with an increase of 15% in female math and vocabulary test scores. We also see significant increases in educational attainment and the schooling of women. Yet, for men, we find much smaller and insignificant effects. The improvement in human capital translates to an about 5% increase in income for females.[4]

Gender preferences are important in China. We present a model of human capital accumulation and parental investments as one potential way to rationalize the empirical results described above. In this model, gender preferences play a part, and parental investments interact with different endowments at birth. The model suggests that with preferences on boys, parents tend to compensate more for boys than girls when they are disadvantaged at birth. The model's predictions are in line with our finding of strong and sizable effects for girls and small and insignificant effects for boys. On the other hand, the model suggests that parental investments can be diverted to different dimensions of skills, such as non-cognitive skills, when there is no need to compensate for the cog-

[4] To calculate the increase in income, we use the results in Wang (2013), who finds that one year of additional schooling raises income with 15% in China.

nitive disadvantage at birth. Similarly, with preferences on boys, the diversion will be stronger for boys than for girls. Indeed, we find that the program has positive effects for boys on non-cognitive skill measures. We find no impact on girls in these non-cognitive skill dimensions. Similar to Dahl, Kotsadam, and Rooth (2017); Dhar, Jain, and Jayachandran (2018); Dossi, Figlio, Giuliano, and Sapienza (2019), we proxy gender preferences by gender attitudes, specifically, about the appropriate roles and rights of women and girls. Across all outcomes, we find that the gains in cognition are most significant for girls born in regions with the strongest son preferences. For boys, we see an opposite heterogeneous treatment effects of iodized salt across districts with different gender attitudes. All the evidence above suggests that gender preferences are an indispensable pathway to explaining the gender difference in the program effects.

Our study contributes to at least four strands in the literature. Firstly, we add to the literature on the long-term effects of early-life conditions. Much of early "fetal origins" work (see, among others, Almond, 2006; Van den Berg, Lindeboom, and Portrait, 2006) has focused on demonstrating the impact of extreme, traumatic experiences (disease outbreak, recessions, famines, severe environmental shocks, etc.) in early life. Recent studies (see Niemesh, 2015; Hoynes, Schanzenbach, and Almond, 2016; Feyrer, Politi, and Weil, 2017; Brown, Kowalski, and Lurie, 2018; Adhvaryu et al., 2020) have shifted the focus to estimating gains to exposure to a purposeful large-scale distribution of resources. The nationally implemented intervention in China started just after the launch of the 1993 WHO campaign. The program is, to the best of our knowledge, the largest of its kind. It is also a commonplace, moderate intervention, and has, therefore, relevant external validity—this aids policymakers in optimizing similar policies in the future. Furthermore, while there exist a large body of studies that look at the impact of in utero exposure to the quantity of food (see Lumey, Stein, and Susser, 2011, for an excellent review of the famine literature), only a few studies (e.g. Field, Robles, and Torero, 2009; Feyrer, Politi, and Weil, 2017; Adhvaryu et al., 2020) have looked at the long-run effects of food *quality* or nutrient intake.

Secondly, and related to the above, we contribute to the discussion on intermediate proxy indicators of long-term outcomes. Adhvaryu, Bednar, Molina, Nguyen, and Nyshadham (2020); Feyrer, Politi, and Weil (2017) examines the long term effect of a salt iodization program, promoted by a private firm, on lifetime income at later ages. Adhvaryu, Bednar, Molina, Nguyen, and Nyshadham (2020) find an about 10% income increase for those exposed to the iodine program. We look at the effect of a public program on childhood cognition and education and find substantial effects. Our study thus adds to a full picture of how early-life disadvantage unfolds over the life course. We use

measurements of human capital that include standardized numeracy tests for all children, as opposed to most of the previous literature, which only focuses on school enrollment. A few recent studies such as (Figlio, Guryan, Karbownik, and Roth, 2014; Almond, Mazumder, and Van Ewijk, 2015; Bharadwaj, Lundborg, and Rooth, 2017; Shah and Steinberg, 2017) take a similar approach as we do by examining the effects of events in early childhood on cognitive test outcomes during the school years. Most studies use administrative data of developed countries, where standardized tests cover most of the school-going children at a certain age. In developing countries, however, a substantial share of the children is already out of school at young ages, and therefore, a similar strategy will only partly measure the effectiveness of the intervention. Moreover, if collected, most human capital measurements are self-reported performance measures, which makes the comparison across individuals difficult. Our work complements the literature with evidence from a vast developing country by using a data set where the results of standardized math and verbal tests are collected for all children, in and out of school.

Thirdly, we also shed light on the literature about child gender preferences. Gender biases favoring males, particularly in education, are more extensive in developing countries like China and India. Females in those countries often receive fewer investments from parents (see, for example, Oster, 2009; Jayachandran and Kuziemko, 2011; Bharadwaj and Lakdawala, 2013; Barcellos, Carvalho, and Lleras-Muney, 2014) and are likely not to reach their full potential in education, health, and personal autonomy. In our current study, the iodine policy has positive (possibly unforeseen) spillovers to females. Our study therefore also speaks to the relevance of early life conditions in explaining gender differences in socioeconomic outcomes later in life.[5]

Finally, our heterogeneous analysis that includes gender preferences is motivated by theoretical models of human capital formation. In such models, often dynamic complementarities between investments at different stages of childhood are considered (Cunha, Heckman, and Schennach, 2010). A handful of studies (Adhvaryu, Molina, Nyshadham, and Tamayo, 2015; Gunnsteinsson, Adhvaryu, Christian, Labrique, Sugimoto, Shamim, and West Jr, 2018; Duque, Rosales, and Sanchez, 2018; Aguilar and Vicarelli, 2018; Rossin-Slater and Wüst, 2018) attempt to identify these dynamic complementarities empirically. For this, they use exogenous variation at different stages of the life cycle and generally cannot find evidence for dynamic complementarities. However, as pointed out by Malamud, Pop-Eleches, and Urquiola (2016), parents might increase investments in the child to counter the adverse effects of the initial shock. This may confound the impact of sub-

[5] See Almond and Currie (2011) who call for work that integrates work on son preference with work on fetal origins.

sequent shocks. They furthermore argue that human capital outcomes for children are the result of parental preferences, the family budget constraint, and the shape of the child health production technology. This makes it challenging to interpret reduced form effect estimates. We use parental gender attitudes as a proxy for parental preferences to identify one of the channels in children's human capital formation and find that gender preferences are important in the formation of the human capital of school-aged children.

Our results point towards four observations that are relevant for the strands of literature referred to above: the relevance of parental investment responses in mitigating the effects of adverse shocks early in life; that child gender preferences are important for these investments decisions; that large scale interventions may crowd out private investments and may also affect skills dimensions initially not aimed at; large scale programs can have positive and possible unintended effects on gender equality in societies where boy preferences are important.

The rest of the paper is organized as follows. Section 3.2 provides a brief overview of Iodine Deficiency Disorders (IDD), the Universal Salt Iodization (USI), and related literature. Section 3.3 describes the data used in the analysis. Section 3.4 outlines empirical model and Section 3.5 discusses the results of the models. We zoom in on our finding of differential effects by gender in section 3.6. This section presents a simple model where gender preferences may differently affect parental investment in girls and boys. We introduce gender attitudes as a proxy for gender preferences and examine whether the program's effects on cognitive and non-cognitive skills vary with gender attitudes. Section 3.7 summarizes our findings, places these findings into context and concludes.

3.2. BACKGROUND

Iodine is an essential component of the hormones produced by the thyroid gland and is therefore essential for human life (Zimmermann, 2011). Insufficient iodine intake causes many disorders from the fetal stage to adulthood, the most common of which is an enlargement of the thyroid gland. Although this enlargement, called goiter, is the most visible symptom of iodine deficiency, besides being inconvenient, it has no severe consequences. However, fetal exposure to iodine deficiency may lead to impaired neurodevelopment. The brain damage caused by severe iodine deficiency in this stage of life is often irreversible.

The knowledge that iodine can help prevent goiter has existed since the mid-1800's (Zimmermann, 2008). It was not until 1895 that iodine was first discovered in the thyroid gland (Baumann, 1896). Switzerland was the first country in the world to introduce

iodized salt in 1922. The United States introduced iodized salt in 1924 after the executive Council of the Michigan State Medical Society officially endorsed iodized salt. In 1993, the World Health Organization (WHO) proposed a worldwide campaign to eradicate IDD. The primary intervention strategy for IDD control is Universal Salt Iodization (USI), a notably simple, universally effective, and particularly cheap instrument. The World Bank reports that it only costs approximately $0.05 per child per year.

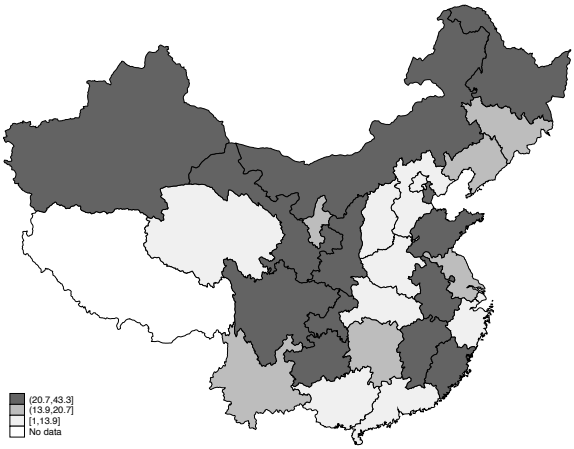
Historically, endemic goiter was found particularly in the mountain regions in China. For instance, in the 1940s, more than 20% of the residents of Kunming, the capital of the province of Yunnan, had goiter (Simoons, 1990). The Chinese Academy of Preventive Medicine had estimated that about 450 million people lived in iodine-deficient areas, with more than 30% of the population considered at risk of IDD (see Chen and Wu, 1998). The iodine deficiency disorders problem was acknowledged as a public health threat, and in response to this in 1993, the State Council of China announced the Universal Salt Iodization (USI) policy to eliminate IDD by 2000.

The Universal Salt Iodization was a national strategy. As the recommended strategy for controlling IDD, the USI requires that all edible salt, including salt for food processing and household use, is iodized. Accordingly, all counties throughout China should supply iodized salt (except for very few (22 out of 1800) officially approved counties). The challenge of the policy was to increase salt iodine levels sufficiently to bring the median urinary iodine concentration of children into the 100–199 ţg/L range and at the same time maintaining the optimal urinary iodine concentration (MUIC) levels of pregnant women (150–249 ţg/L). To reach the desired intake of iodine, the State Council enacted in October 1994 the national regulation of salt iodization. The level of salt iodization during the manufacturing process was set at 50 mg/kg in 1994.

Besides, between 1993 and 1995, a national monitoring system was built to track trends in goiter prevalence among school children aged 8-10. The monitoring was held between March and June 1995. In our empirical analyses (see section 3.4), we will use the outcome of this monitoring exercise as the pre-policy distribution of iodine deficiency levels across the different provinces. Note that this is a few months after the implementation of the Salt Iodization Program. Therefore, a concern may be that the cross-province variation in goiter rates does not reflect the pre-policy distribution of iodine deficiency rates. It is important to stress that the literature has documented lags of at least one year before goiter rates normalize after iodine repletion (Pardede, Hardjowasito, Gross, Dillon, Totoprajogo, Yosoprawoto, Waskito, and Untoro, 1998; Jooste, Weight, and Lombard, 2000; Zimmermann, Hess, Adou, Toresanni, Wegmüller, and Hurrell, 2003).[6]

[6] Pardede, Hardjowasito, Gross, Dillon, Totoprajogo, Yosoprawoto, Waskito, and Untoro (1998) and Jooste,

Figure 3.1: Goiter Distribution in 1995



Notes: Figure 3.1 reports goiter rates (%) (among schoolchildren aged 8-10) in 1995. Darker areas represent higher goiter rates.
Sources: National Iodine Survey 1995

Figure 3.1 shows the pre-policy spatial distribution of iodine deficiency levels of schoolchildren aged 8-10. Pre-policy (1995) goiter rates among children under ten years old do not differ significantly by gender (Sun, 2018). The dark areas (mostly western and northern provinces) indicate high prevalence rates (up to 43.3%), while the light areas (southeast) indicate low prevalence rates. For our empirical analyses, it is of importance to know whether the Universal Salt Iodization policy was effective in increasing the urinary iodine concentration levels in the population. We turn to this below in the data section.

3.3. DATA

3.3.1. GOITER DATA

The base, pre-policy, geographic distribution of goiter prevalence before the salt iodization policy (see Figure 3.1) came from the 1995 National Iodine Survey on goiter rates[7] among schoolchildren.[8] In each provincial survey, a multistage, probability proportional

Weight, and Lombard (2000) documented non-significant reductions in the size of the thyroid gland among children age 12 in South Africa one year after the introduction of iodized salt. Zimmermann, Hess, Adou, Toresanni, Wegmüller, and Hurrell (2003) found in children aged 8-9 in Cote d'Ivoire even two years after the salt intervention only an eight percent reduction in the goiter rates.

[7] The goiter rate is defined as the percentage of schoolchildren who have either Class I or Class II goiter. Class I goiter in normal posture of the head cannot be seen, and it is only found by palpation. Class II goiter is palpable and can be easily seen. Goiter rates in previous studies are the prevalence of Class II goiter as the information mostly comes from historical data.

[8] Although there was no representative measures of goiter prevalence before 1995, goiter prevalence of 15 areas (from 10 provinces) were measured in three consecutive years (1991, 1992, and 1993). The average goiter rate

to the population size cluster sample was obtained. The county served as the primary sampling unit, and in each province, 30 counties (clusters) were selected from a county population list. In each selected county, a school was then sampled at random. Children aged 8 to 10 years at the time of the survey served as the index population. For each cluster, 40 children were selected at random from the enrollment list. All children were examined for thyroid size by palpation.[9] Therefore, the goiter rate is defined as the percentage of schoolchildren who have either Class I or Class II goiter. The sample sizes by province ranged from 1,200 to 2,400 (mean, 1,259). Our goiter data has an important advantage over goiter measures used in some recent studies like Feyrer, Politi, and Weil (2017); Adhvaryu, Bednar, Molina, Nguyen, and Nyshadham (2020) who use goiter prevalence among military recruits. This index population consists of young and healthy and as such may not be a representative measure of local iodine deficiency problems.

The survey was held every two or three years, enabling us to track the effectiveness of the Universal Salt Iodization program over time. Indeed, the program was proved to be very effective. By 2002, provinces converged to very low child goiter rates, so that provinces with high pre-eradication levels of goiter experienced the largest reductions. This is illustrated in Figure 3.2a that shows average goiter rates (%) across China over 1995-2002. The average goiter rate decreased from 20% in 1995 to around 5% in 2005. Figure 3.2b shows the post-campaign decline in goiter rate versus pre-campaign levels. Of importance for our empirical analyses is that this figure shows that the policy was effective in bringing down goiter rates for all provinces.

3.3.2. THE SAMPLE, OUTCOME VARIABLES AND CONTROL VARIABLES

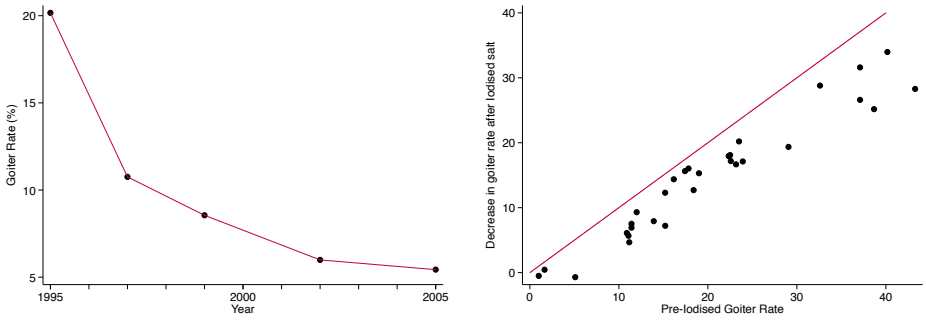
THE SAMPLE

The micro-level data used in this study come from the *China Family Panel Studies* (CFPS). The CFPS is a large-scale nationally representative panel survey conducted by the Social Science Survey Institute at Peking University. Three waves of the survey have been published until 2017. The CFPS baseline wave (hereafter CFPS-2010) selected a total of 14,798 households, containing 33,600 adults and 8,990 children. A second and third wave of the CFPS followed the same individual from the households in 2012 and 2014 (hereafter CFPS-2012 and CFPS-2014). A standard math and verbal tests were carried out in CFPS-2010 and CFPS-2014. For all three waves, educational attainment was recorded. Besides cognitive measures also, some non-cognitive measures were collected. We provide more information on these measures in section 3.6. To maintain the consistence of

is 12.9, 15.9, and 13.6 in these 15 areas, which supports that the decline of goiter prevalence started from 1995.

[9] In some provinces, part of the children were also examined by ultrasound.

Figure 3.2: Goiter Prevalence Before and After the Intervention



(a) Goiter Rates Decline (b) Before and After the Intervention

Notes: Figure 3.2a reports time-series data on mean goiter rate (among schoolchildren aged 8-10) across country between 1995 to 2005. Figure 3.2b shows the post-intervention decline in goiter rate versus pre-intervention levels across China.
Sources: National Iodine Survey 1995-2005

the sample selection, we pool the baseline wave (CFPS-2010) and the third wave (CFPS-2014) together. The data include accurate birth information, such as year, month of birth, place of birth, and whether individuals were born in a rural area. We restrict ourselves to those born in the rural areas (81% of the total population). The intervention is likely to be cleaner for those born in a rural area, as in the urban areas people had better access to micronutrient food supplements.[10] The survey also collects respondents' migration history. Migration at young ages is very low (less than 3%). The Salt Iodization policy was implemented in October 1994. We include cohorts born between July 1990 and June 2000. At the base wave (CFPS-2010), these individuals were between 10 and 19 years old. This leaves us with 4039 children at the baseline survey. For 4021 children individuals, we have all the key information on test scores and education. The third wave (CFPS-2014) included the same math and verbal tests as the baseline wave. We, therefore, added this third wave to our baseline data. This resulted in 441 additional children. However, due to sample attrition between 2010 and 2014, we also lost 1069 individuals.[11] In the end, we have 7414 observations on schooling (3856 males and 3558 females) and 6420 observations on test scores (3310 male and 3110 female test scores). In the analyses, we link these individual observations to the pre-policy goiter rates.

[10] The use of such supplements are also likely to vary by parental Socio-Economic Status (SES), which in itself correlates strongly with our cognitive and educational outcomes measures.

[11] In section 3.5.3, we show that the main findings are robust to the sample attrition by only using the data in the baseline.

VARIABLES

Educational attainment is an ordered categorical variable ranging from one to four (illiterate, graduated from primary school, middle school, and high school). In the regression analysis, we use a dummy variable, indicating whether individuals have graduated from primary school. The number of years in school measures schooling. As alternative measures of human capital, we also use a math test and a verbal test designed by the CFPS team. Both tests were presented to the students, irrespective of their age. The math test was designed to test for primary and secondary school math knowledge and consists of twenty-four mathematical problems. Questions were sorted in order of increasing difficulty, and each of those questions counted for one point. Similarly, the verbal test consists of thirty-four Chinese characters based on the language textbooks again. Characters, which counts for one point each, are sorted in order of increasing difficulty. Therefore, full scores of the math and language test are 24 and 34, respectively. Note that as we measure cognition and schooling at fixed points in time, older children will perform better. It will, therefore, be important to control for age in the later regressions. Our primary empirical strategy controls for age in a very flexible way by fully exploiting the panel structure of the data set. Particularly, we always control for CFPS wave by age (dummies) interactions. Additionally, CFPS waves by cohort interactions are included to make sure we are always comparing individuals with same age. The sample also includes basic socio-demographic variables, such as age, gender, parental educations, birth order, and family size.[12] In our empirical strategy, it will be important to control for the possible mean reversion. Therefore, we supplement the set of individual controls with a series of province-level pre-policy characteristics. All these variables are listed in Appendix A, where we also describe how variables are constructed.

Table 3.1 reports summary statistics of socio-demographic variables by gender for provinces with high initial (pre-policy) goiter rates (goiter rates above the median of 17%) and low initial goiter rates (goiter rate prevalence less than 17%) for cohorts born before the implementation of the program (1991-1994).[13] The Table shows that parental education is higher in regions with low initial goiter rates. The table also shows that female outcomes on test scores and education are slightly better than the outcomes for males. However, the table also shows that the differences in schooling and test scores between high and low goiter regions are larger for girls than for boys.

[12] Families in the rural area were regularly exempted from the one-child policy. For example, rural married couples were allowed to have a second child if the first child was female (Zhang, 2017). The sex ratio in our sample of rural born children is about 1.07, which is lower than 1.12 (i.e., the sex ratio at birth in the 1990s documented in Jayachandran (2015)).

[13] In the Appendix, we provide results for the cohorts born after the implementation of the salt iodization policy (Table A2).

Table 3.1: Summary Statistics

	High Goiter Provinces		Low Goiter Provinces	
	Females	Males	Females	Males
Outcomes				
Educational Attainment	2.96	2.93	3.13	3.03
	[0.85]	[0.86]	[0.76]	[0.78]
Illiterate	0.053	0.066	0.016	0.0090
Primary School	0.22	0.21	0.18	0.26
Middle School	0.43	0.45	0.46	0.43
High School or above	0.29	0.27	0.34	0.31
Schooling	9.95	9.78	10.5	10.2
	[3.11]	[3.23]	[2.58]	[2.59]
Math Test Scores	15.3	15.2	16.3	15.6
	[5.67]	[5.89]	[4.90]	[5.39]
Verbal Test Scores	26.3	25.2	27.0	25.3
	[7.23]	[7.75]	[5.94]	[7.17]
Demographics				
Age	19.1	19.3	19.4	19.4
	[2.58]	[2.58]	[2.55]	[2.62]
Father's Educational Attainment	2.29	2.26	2.52	2.56
	[0.96]	[0.99]	[0.93]	[0.90]
Mother's Educational Attainment	1.70	1.72	2.10	2.04
	[0.87]	[0.87]	[0.92]	[0.90]
Birth Order	1.59	1.57	1.61	1.79
	[0.83]	[0.79]	[0.82]	[0.94]
Family Size	4.89	4.57	4.89	4.67
	[1.47]	[1.44]	[1.42]	[1.52]
Number of observations	810	895	795	926

Notes: Author's tabulations of CFPS-2010 and CFPS-2014. Sample consists individuals born in rural area between July 1990 and June 1995. We label a province as high/low goiter if its goiter rate is above/below 17% (median).

3.4. EMPIRICAL STRATEGY

3.4.1. BASELINE ECONOMETRIC MODEL

As we discussed in the previous section, salt iodization was rolled out nationwide in October 1994. Therefore, no province could serve as a pure control group. As a proxy for pre-policy iodine deficiency rates, we use province goiter rates among 8-10 years old children at the start of 1995 (see Figure 3.1). Like, among others, Bleakley (2010a); Adhvaryu, Bednar, Molina, Nguyen, and Nyshadham (2020), we use a difference-in-differences design. In this way, we compare trends in various outcome measures in provinces with different levels of iodine deficiency before implementing the salt iodization program.

We define someone as treated if the entire gestation period[14] is after the date of the

[14] As is generally done in the literature, we use a nine-month gestation period.

implementation of the salt iodization program. All others are considered to be controls. In section 3.5, as a robustness check, we also consider alternative definitions to assign treated and controls in groups. In contrast to earlier studies that used the 1924 salt iodization policy in the U.S., the Chinese iodized salt campaign was implemented rapidly across the entire country. At the start of 1995, more than 80% of the families had already access to iodized salt. Two years later, this has increased to 95%.

During the gestational period, iodine primarily affects the fetus' neurodevelopment, with consequences for postnatal cognitive functioning (Cao, Jiang, Dou, Rakeman, Zhang, O'donnell, Ma, Amette, DeLong, and DeLong, 1994; Zimmermann, 2011). Therefore, we focus in the first instance on the impact of the salt iodization policy on cognitive ability and school attainment of children who are affected in utero. For this, we use the following baseline regression:

$$Y_{ipt} = \beta_0 + \beta_1 Post_t \times Goiter_p + X_{ipt}\rho + \delta_p + \gamma_t + \epsilon_{ipt} \tag{3.1}$$

where outcome Y_{ipt} is either the logarithm of the cognitive test scores (math and verbal), educational attainment as well as the number of schooling years for individual i, who was born in province p, in year t. $Post_t$ indicates whether the individual was conceived after the introduction of iodized salt. $Goiter_p$ is a measure of pre-eradication endemicity in individual i's province of birth. As mentioned above, we use goiter rates collected in the National Iodine Survey held at the start of 1995. The mean goiter rate is 20%, and the standard deviation is 12%.

The vector X_{ipt} includes individual characteristics: parents education and family size, and mean-reversion controls.[15] First, we follow Bleakley (2010b) and construct the mean-reversion control by interacting provincial average educational attainment in the 1990 Census with the dummy variable $Post_t$. Second, we control for many baseline province pre-treatment characteristics interacted with cohort dummies. In our baseline model, we control for hospitals per capita in 1991, hospital beds per capita in 1991, and the sex ratio in Census1990, again all interacted with cohort dummies. In a robustness check (section 3.5.3), we additionally control for the number of schools per capita in 1991, poverty rates in 1993, and average household income in 1991 all interacted with cohort dummies.[16]

To take into account age effects and the year of survey effects, CFPS wave by age

[15] If the oldest cohorts had high Iodine Deficiency Disorders and low human capital because of some mean-reverting shock. We might expect human capital gains for the subsequent cohorts, even in the absence of a direct effect of the salt iodization policy on productivity.

[16] The provincial poverty rates in 1993 are obtained from Woo, Li, Yue, Wu, and Xu (2004). The average household income in 1991 is constructed from the National Fixed Point Survey.

(dummies) interactions are also included. Additionally, CFPS waves by cohort interactions are included to account for differential trends in the outcome variables. δ_p and γ_t are province and birth cohort fixed effects. δ_p and γ_t are province and birth cohort fixed effects. Note that the birth cohort fixed effects are important here as our outcome variables are measured in 2010 and 2014, which implies that those exposed to the salt iodization are much younger than the controls. We also control for region-specific linear trends in all models.[17] We run specification (3.1) separately for males and females. We also consider alternative specifications in section 3.5.3. Of prime interest is the continuous treatment variable $Post_t \times Goiter_p$ that proxies potential iodine exposure. Recall from Figure 3.2b that the salt iodization policy was very effective in reducing goiter rates in all provinces to very low levels. So while the parameter β_1 in a strict sense is the intention-to-treat effect, the high compliance rates make it very close to the treatment effect.

3.4.2. DYNAMIC SPECIFICATION

Since our estimates use the cross-province convergence in goiter rates created by the introduction of iodized salt (Figure 3.2a and 3.2b), convergent pre-trends across high and low-base goiter rate provinces prior to 1995 are a concern. Therefore, we also use an event study design to test for the common pre-trends assumption formally. More specifically, we run the following regression:

$$Y_{ipt} = \beta_0 + \sum_{t=1990}^{2000} \beta_t \times Goiter_p + X_{ipt}\rho + \delta_p + \gamma_t + \epsilon_{ipt}, \tag{3.2}$$

where β_t gives the cohort-specific relationship between pre-eradication endemicity and later-childhood outcomes.[18] If salt iodization affected the human capital formation of exposed cohorts, these effects should be visible in a break from pre-existing trends in β_t. This method will also shed light on the partial effects of iodine exposure in late childhood (rather than in utero) if such effects exist.[19] Note that all individuals born in 1995 or later are exposed to iodized salt from conception onward. Individuals born in 1994 experience higher iodine intake in their year of birth; thus, this cohort is partially exposed

[17] The regions consist of several provinces. See Table A1 of the Appendix A for the precise definition.

[18] In practice, β_{1994} represents estimates of individuals born between July 1994 and June 1995, and β_{1995} for the cohort born between July 1995 and June 1996, etc. We made such adjustments because individuals born after July 1995 were conceived after implementing the salt iodization policy. Similarly, the 1993 cohort are those born between July 1993 and June 1994.

[19] See Zimmermann (2011) for a comprehensive summary of the role of iodine in human growth at different stages in life. For example, neonatal iodine deficiency may cause endemic cretinism. Deficiency during childhood and adolescence may impair mental functioning and delay physical development.

to higher iodine levels in utero and fully exposed from birth onward. Individuals born in 1993 experience higher iodine from age one onward, and those born in 1992 experience higher iodine from age two onward. Since we normalize the 1994 cohort coefficient to zero, our analysis essentially tests for differential effects of exposure relative to exposure at age one and older. If there are additional benefits to having access to iodine between conception and age 1, we would expect the coefficients β_{1995+} to be positive. Similarly, if iodine at age 1 has an additional benefit relative to iodine exposure at age two or older, we would expect coefficients β_{1993-} to be negative.

3.5. RESULTS

3.5.1. BASELINE RESULTS

Table 3.2 reports the main results of two separate regressions of our basic model (Equation 3.1): one for men in Panel A and the other for women in Panel B. In all the regressions discussed in this section, the coefficients of interest are the post-by-goiter rate interaction, which represent the effect of salt iodization on our outcomes of interest. The coefficients in all tables have been multiplied by 12, the inter-quartile (25-75) range of the goiter distribution. The size of the effect is scaled to be the effect of moving from a relatively high goiter province to a low goiter province. Although the following tables only report the coefficient of interest, in all specifications, we include controls for province and year of birth fixed effects, birth order, family size, parents' characteristics, region-specific linear trends, a series of mean reversion controls, age and the survey year controls. Standard errors are clustered at the province-of-birth level to allow for arbitrary correlation of the errors for individuals born in the same province. We also report two-way clustered standard errors by province and family (in parenthesis) to take account of situations where there are multiple children in a family. The standard errors of two-way clustering are almost identical to the usual cluster robust standard errors. Given that our data only contain 28 provinces, we produce statistical inference based on the wild-bootstrap approach (Cameron, Gelbach, and Miller, 2008). The associated wild-bootstrap standard errors (in angle brackets) turn out to be similar to the usual cluster robust standard errors. For all four human capital measures, we identify significant effects of the intervention for females. A one standard deviation decrease in the pre-intervention goiter rate is associated with about 4% increase in the probability of graduating from primary school, a 10% increase in schooling, 14% increase in math and 12% increase in verbal test scores. For males, the coefficients are substantially smaller in magnitude and not significant for any of the outcome variables.

Table 3.2: Iodine Exposure and Human Capital Attainment

	(1)	(2)	(3)	(4)
	Math Test ln(scores)	Verbal Test ln(scores)	Primary School	Schooling ln(years)
Panel A: Males				
Post × Goiter	0.0298	0.0605	0.0171	0.0228
	[0.0508]	[0.0511]	[0.0226]	[0.0411]
	(0.0566)	(0.0666)	(0.0234)	(0.0370)
	⟨0.0553⟩	⟨0.0492⟩	⟨0.0210⟩	⟨0.0428⟩
Mean of Dep. Var.	2.517	3.083	0.497	2.108
Observations	3310	3310	3791	3654
Panel B: Females				
Post × Goiter	0.137	0.117	0.0382	0.108
	[0.0372]***	[0.0556]**	[0.00811]***	[0.0202]***
	(0.0310)***	(0.0468)**	(0.0134)***	(0.0185)***
	⟨0.0312⟩***	⟨0.0507⟩**	⟨0.0122⟩***	⟨0.0152⟩***
Mean of Dep. Var.	2.543	3.170	0.528	2.123
Observations	3110	3110	3510	3386

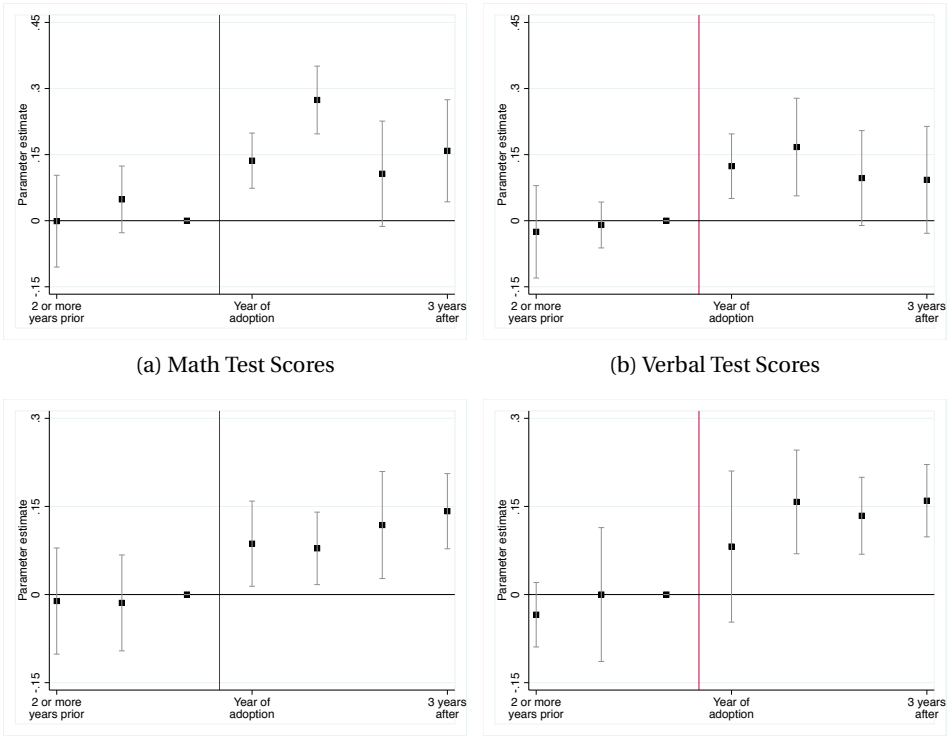
Notes: Each coefficient is from a separate regression. All regressions control for fixed effects specific to birth province and birth year, birth order, family size, parents' education, region-specific linear trends, survey wave by age interactions and survey wave by cohort interactions. Mean-reversion controls include provincial average educational attainment in the 1990 Census interacted with the dummy for treated cohorts, hospitals per capita in 1991, hospital beds per capita in 1991, and the sex ratio in Census1990 all interacted with cohort dummies. Standard errors clustered by province appear in square brackets. Two-way clustered standard errors by province and family appear in parenthesis. Standard errors based on wild-bootstrap approach (Cameron, Gelbach, and Miller, 2008) with 999 replications appear in angle brackets. *, **, *** indicates significance at the 10%, 5% and 1% level respectively.

3.5.2. RESULTS FROM THE EVENT STUDY

Crucial for the causal interpretation of our coefficients is the assumption of common pre-intervention province trends in the outcome variables. In order to test for differential pre-trends, we run regressions of the form specified in Equation (3.2). The Figure 3.3 plots the estimated coefficients (β_t) of the event study for females. Results for males are reported in Appendix B (Figure B1). The figure shows that the trends leading up to the year of the intervention are identical and insignificant from the 1994 effect. This gives confidence in the validity of the common trend assumption. Figures 3.3a and 3.3b show that math and verbal scores of females increase shortly after the intervention with about 10%. For educational outcomes (Figure 3.3c and 3.3d), we see a similar picture (about 10% for both educational attainment and years of schooling, respectively). While there is some variation in the effect size of the program across the cohorts, the 95 percent confidence intervals overlap for all cohorts. This suggests that females born in the later cohorts obtain comparable benefits from the adoption of iodized salt. The results

for males in Appendix B point at insignificant effects for all outcome variables and all cohorts.

Figure 3.3: Event Study



(a) Math Test Scores

(b) Verbal Test Scores

(c) Primary School

(d) Schooling

Notes: The sample includes all female respondents from two waves of the survey (CFPS-2010 and CFPS-2014). Each point reflects the coefficient estimated on an interaction term between the birth year (compared to 1995) and the pre-intervention (base) level of the goiter rate in the birth-province. Capped spikes represent 95 percent confidence intervals. All models condition upon birth province and birth year fixed effects and the full set of controls used in our main analysis.

3.5.3. ROBUSTNESS ANALYSIS

There are several reasons why trends in educational outcomes across birth cohorts might differ across provinces. In this section, we explore some additional specification checks to make sure that our baseline results can be interpreted as causal. Table 3.3 presents these additional specifications/checks. Here we restrict ourselves to the results of robustness checks for females. The same robustness checks for males are reported in Table B1 of Appendix B.

Table 3.3: Robustness Checks (Female)

	(1)	(2)	(3)	(4)
	Math Test ln(scores)	Verbal Test ln(scores)	Primary School	Schooling ln(years)
Panel A: raw test scores				
Post × Goiter	0.814	0.667	0.0399	0.493
	[0.239]***	[0.439]	[0.00783]***	[0.108]***
Mean of Dep. Var.	13.90	25.23	0.530	8.730
Observations	3110	3110	3522	3386
Panel B: additional controls				
Post × Goiter	0.108	0.106	0.0230	0.0842
	[0.0437]**	[0.0509]**	[0.0224]	[0.0346]**
Mean of Dep. Var.	2.543	3.170	0.530	2.123
Observations	3110	3110	3522	3386
Panel C: drop partial exposed group				
Post × Goiter	0.124	0.113	0.0386	0.106
	[0.0395]***	[0.0536]**	[0.00754]***	[0.0183]***
Mean of Dep. Var.	2.540	3.169	0.529	2.120
Observations	3026	3026	3426	3297
Panel D: only using baseline wave 2010				
Post × Goiter	0.129	0.105	0.0442	0.100
	[0.0310]***	[0.0468]**	[0.0142]***	[0.0185]***
Mean of Dep. Var.	2.491	3.128	0.343	1.975
Observations	1861	1861	1906	1906
Panel E: small sample window				
Post × Goiter	0.137	0.136	0.0322	0.0899
	[0.0278]***	[0.0495]**	[0.0106]***	[0.0201]***
Mean of Dep. Var.	2.580	3.196	0.553	2.146
Observations	2529	2529	2877	2755

Notes: Each coefficient is from a separate regression. All regressions except Panel B use the same controls as the baseline model in Table 3.2. In Panel B, we control for birth-region and birth-year specific interaction instead, and we additionally control for schools per capital in 1991, poverty rates in 1993, and average household income in 1991 all interacted with cohort dummies. Standard errors clustered by province appear in square brackets. *, **, *** indicates significance at the 10%, 5% and 1% level respectively.
Data: CFPS-2010 and CFPS-2014

To ensure that the logarithm transformation does not drive our finding, we run the same regression using raw math and verbal scores as dependent variables. The results reported in Panel A show that the conclusions from the baseline models remain to hold and that the functional specification does not drive the findings.

Another concern of our identification strategy is mean reversion. In the baseline, we control for many baseline province pre-treatment characteristics interacted with co-

hort dummies. To further decrease the concern for mean reversion, we also control the number of schools per capita in 1991, poverty rates in 1993, and average household income in 1991 all interacted with cohort dummies. The provincial poverty rates in 1993 are obtained from Woo, Li, Yue, Wu, and Xu (2004). The average household income in 1991 is constructed from the National Fixed Point Survey. Instead of region-specific linear trends, we control for the region of birth by birth year interactions. The results of this exercise, reported in Panel B, show that the effects remain significant and are quantitatively similar (if not, larger) than the baseline estimates. A one standard deviation decrease in the pre-intervention goiter rate is associated with a roughly 15% increase in cognitive ability and 11-13% in educational outcomes for females.

In section 3.3, we assigned individuals into the treatment group if the full gestation period was after October 1994. All others, the partially exposed and those who were not exposed, were placed in the control group. This assumes that the gestational period was precisely nine months and that those whose gestational period partly lies after October 1994 are not exposed in utero. This classification is perhaps a bit too conservative. Panel C reports the results using the sub-sample where the partially exposed children (born in April, May, and June 1995) are dropped from the analyses. The results in Panel C show that this hardly affects the estimates.

The verbal and math test remained the same over the 2010 and 2014 wave. In the basic specification, we pooled the observations from the 2010 and the 2014 wave. Individuals in our analysis are in their late teen ages/early-adulthood, and therefore as we do in the regressions controlling for age is important. However, using information from all waves has a threat that individuals might learn from the test results in the first wave. This makes the second wave test scores inappropriate as a measurement of development in cognition. As a check, we run regressions using only the data from the first wave of the CFPS. Results of using only first wave in 2010 is also informative on the impact of sample attrition between the first wave in 2010 and the last wave in 2014. Indeed, we lost around 30% percent of observations due to attrition (see section 3.3). The results in Panel D show that the results remain close to our baseline specification and confirm our main findings.

We also used a slightly smaller sample window by restricting individuals born between July 1991 and June 1999. The advantage is a lower attrition rate in wave 2014 as older cohorts (born in 1990) are more likely to have left home. Moreover, the trimming of the left and right tail of the age distribution makes the sample more homogeneous. A downside of this age/cohort restriction is a smaller sample size and hence less power. Still, however, all four estimates in Panel E remain to be significant and close to the base-

line estimates in Table 3.2.

We also performed a placebo test, where use parental education (which was used as a control variable in the main specifications) as outcome variables. The idea of this test is that parental education should not be affected by future exposure to iodized salt, and if we do find an effect, it may suggest that our results may be driven by parental background. Indeed, both mother and father's education (measured by education attainment or years of schooling) are not affected by future exposure (results available upon request).

Finally, we further verify our identification assumptions and demonstrate the statistical power of our inferences by conducting falsification tests where we assign a pseudo-treatment. More precisely, we randomly assign province of birth and thus pre-intervention goiter rates to each respondent in our sample. If our identification strategy is valid, we would expect estimates using those pseudo-samples to be centered around zero. We can then confront our baseline estimates with the results from the pseudo-sample. In the Figure 3.4, we plot the distribution of the t-statistics from 5,000 estimated pseudo-treatment effects on educational attainments, schooling, math and verbal ability, respectively. As expected, all four distributions are centered around zero. Together, these results imply that assumptions in our empirical model are unlikely to be violated. To address the statistical power of our model, we mark within the distribution of pseudo-treatment effects the location of the t-statistic of the baseline treatment effects in Table 3.2. We also report at the top of each figure the share of the pseudo-treatment t-statistics that exceed the actual t-statistic of the baseline model (in absolute values).[20] These p-values give confidence in our design and statistical power of our exercises.

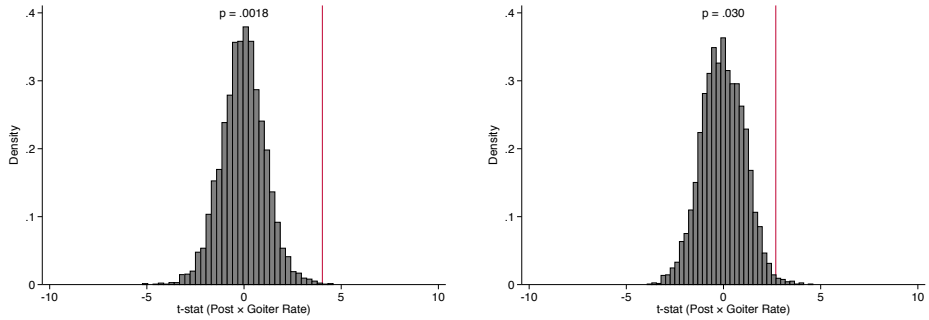
3.5.4. ENDOGENOUS SEX SELECTION

In this section, we address the potential threat posed by endogenous sex selection. Goiter rates may influence the gender ratio for biological reasons. Alternatively, there could be parental motives related to goiter prevalence rates to sex select. In the presence of ultrasound techniques, such selection effects are not inconceivable. In either case, this would imply that the marginal girl born in a high goiter area would be different from the marginal girl born in a low goiter area, which may pose a threat to the interpretation of our findings.

To rule this out, we test directly whether the provincial sex ratio is correlated with iodine deficiency (goiter prevalence rates). More importantly, we test whether the associa-

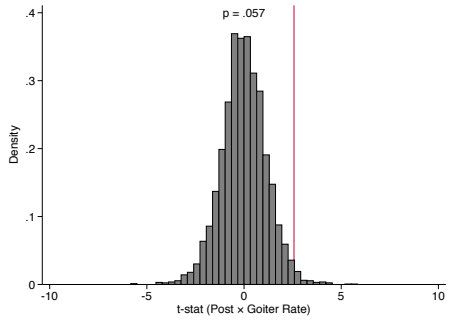
[20] These p-values can be seen as alternatives to the p-values obtained from our clustered standard errors reported in Table 3.2.

Figure 3.4: Falsification Tests

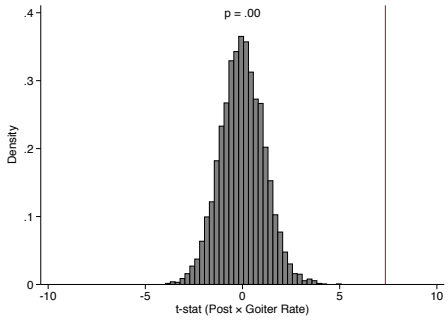


(a) Math Test Scores

(b) Verbal Test Scores



(c) Primary School



(d) Schooling

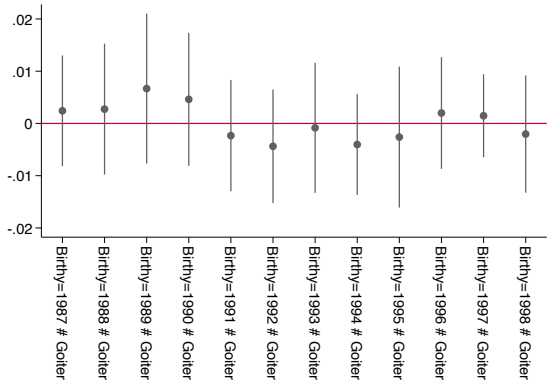
Notes: Pseudo-treatment vs. actual policy intervention: the distribution of t-statistics resulting from 5,000 random assignments of treatment to individuals, as well as the t-statistics from the actual treatment through the policy intervention (red line). "p-values" report the share of the pseudo-treatment t-statistics that is larger than the actual t-statistics.

tion changes after 1994. Following Edlund, Li, Yi, and Zhang (2013) we use the Census of 2000 to calculate the sex ratio by the birth year (1987 to 1999) at the province level. Linking this data with provincial goiter rates, we display in Figure 3.5 the results of an event study regression that relates provincial sex ratios to goiter prevalence rates. As may be clear from the graph, there is no association between the sex ratio and iodine deficiency, and the association does not change after introducing the salt iodization policy.

3.5.5. COMPARISON WITH OTHER COHORT-BASED IODINE STUDIES

Some recent studies (Field, Robles, and Torero, 2009; Feyrer, Politi, and Weil, 2017; Adhvaryu, Bednar, Molina, Nguyen, and Nyshadham, 2020) also analyze the long-run impacts of iodine deficiency in early-life. Field, Robles, and Torero (2009) was one of the

Figure 3.5: Sex Ratio and Iodine Deficiency



Notes: Sex ratios are aggregated from microdata of the 2000 Census. Following Edlund, Li, Yi, and Zhang (2013), we calculate the sex ratio at province level for each birth year between 1987 and 1999. In the event study regression, we control for province fixed effects and year of birth fixed effect. Therefore, cohort who were born 1999 is the baseline group and the coefficient of the cohort 1999 is then omitted in the graph. Standard errors are clustered by province.

first to provide evidence at the micro-level of the effect of iodine availability in utero on schooling attainment in Tanzania. Feyrer, Politi, and Weil (2017); Adhvaryu, Bednar, Molina, Nguyen, and Nyshadham (2020) exploit a nationwide salt-iodization program initiated by the public health authorities in Michigan in 1924.[21] Similar to our study, they use pre-program geographical information on goiter prevalence along with the time variation in the introduction of iodized salt to assess the causal effect of iodine on later life outcomes. Feyrer, Politi, and Weil (2017) find strong effects for males: iodized salt in utero leads to a 15 point increase in I.Q. Interestingly, Adhvaryu, Bednar, Molina, Nguyen, and Nyshadham (2020), using census data, find no effects for males, but strong increases in income (11%), labor force participation (0.68%) and full-time work (0.9%) for females. Field, Robles, and Torero (2009) also find stronger effects for girls. Like Adhvaryu, Bednar, Molina, Nguyen, and Nyshadham (2020), we only find effects for females. Relating our findings to Field, Robles, and Torero (2009) and Adhvaryu, Bednar, Molina, Nguyen, and Nyshadham (2020), we find that the introduction of iodized salt leads to 0.441 additional years of schooling (see panel A of Table 3.3). This estimate is in line with Field, Robles, and Torero (2009), who find that the iodine supplement program increased schooling years with 0.35-0.56 years. Wang (2013) finds for China that one year of additional schooling raises income with 15%, which translates to an about

[21] Morton's salt, the largest producer in the U.S. at that time, began selling iodized salt in the fall of 1924.

5% (0.441 × 0.11) increase in income for females in our study.

Our strong effects of iodine in utero for females is consistent with the medical literature that posits that female fetuses are more sensitive to maternal thyroid deficiency than male fetuses Zimmermann (2011). However, this does not imply the absence of effects for males (see, for instance, the findings of Feyrer, Politi, and Weil (2017) and Field, Robles, and Torero (2009)). In developing countries, gender differences in socioeconomic outcomes are bigger than in developed countries, and cultural gender norms may contribute to gender differences in human capital outcomes. These gender norms may already have an impact on children very early in life and might explain heterogeneous gender effects found in the literature on long-run effects of early life shocks.[22] Indeed, parents' preferences for boys make these parents to desire more sons than daughters, but it may also result in parents choosing to invest more in sons than in daughters (Jayachandran, 2015). The former mechanisms may result in a male-skewed sex ratio. Almond and Edlund (2008); Abrevaya (2009) find that son preferences in Asian immigrants to the U.S. seem to persist with changing the economic environment, suggesting a strong role for culture (i.e., preferences). Important for our study is that son preferences may imply not only more material and non-material resources in boys post-birth but also that parental investments post-birth may mitigate some of the adverse effects of iodine deficiency in boys more than in girls. This would be consistent with our finding of the strong effects of the salt iodization policy for girls and no effect on boys. Below we discuss the role of parental investments, which helps to explain the mixed finding in the literature concerning gender.

3.6. MORE ON GENDER DIFFERENCES

3.6.1. CONCEPTUAL FRAMEWORK

To explicitly consider possible interactions between gender preferences in receiving parental investment and endowment early in life, we consider a slightly adjusted version of the model of child human capital formation (Yi, Heckman, Zhang, and Conti, 2015; Almond, Currie, and Duque, 2018). In the model, each child has two components of human capital: cognitive skills (θ^C) and non-cognitive skills (θ^N). Parents can influence the formation of human capital by making investments in the child's skill dimensions I^k, $k \in \{C, N\}$. The optimal investment decision follows from a maximization of the parents'

[22] This was recently also argued by Dinkelman (2017), who refers to the effects of local shocks to the environment in the Asian context that affect resource availability. Findings for Indonesia (Maccini and Yang, 2009) and India (Pathania, 2007) point at more negative effects for girls.

utility function:

$$U = U(c,q) = (1-\alpha_s)\log c + \alpha_s \log q(\theta^N, \theta^C), \tag{3.3}$$

subject to a budget constraint and the child's human capital production function, where $s \in \{boys, girls\}$, c parental consumption and q child quality. The parameter α_s captures the preferences for child quality and may vary across families. Son preferences imply higher values of α_s for boys, $\alpha_{boys} > \alpha_{girls}$. From the optimization of the model, it follows that the optimal investment depends on the child's initial endowments, the price of consumption, the price of investments in skills and preferences α_s. Note that investments could be broadly interpreted. It could be well-targeted efforts aimed at improving specific skills as well as the provision of an environment (e.g., food, housing, attention, etc.) that foster child well-being and child outcomes. In this multidimensional model of human capital formation, investment strategies of parents depend not only on parental preferences but also on the production technology and the budget constraint.[23]

Following Yi, Heckman, Zhang, and Conti (2015); Almond, Currie, and Duque (2018), the total effect of a shock early in life (e) on cognition can be decomposed into two parts:

$$\underbrace{\frac{\mathrm{d}\theta^C}{\mathrm{d}e}}_{A} = \underbrace{\frac{\partial\theta^C}{\partial e}}_{B} + \underbrace{\frac{\partial\theta^C}{\partial I^C}}_{C} \times \underbrace{\frac{\partial I^C}{\partial e}}_{D}. \tag{3.4}$$

A similar expression can be obtained for non-cognitive outcomes (θ^N). The term (A) on the left-hand side of (3.4) is the total effect of an early-life shock and corresponds to the usual reduced form estimates in the empirical literature. In our study, before the salt-iodization policy, the term (A) would be the effect of iodine deficiency, i.e., the pre-intervention cognitive outcomes in late childhood due to geographical variation in iodine levels. The first term on the right-hand side (B) is the biological effect that directly operates through the production function. The second term ($C \times D$) is a behavioral effect from the parental investment response, where (C) is the productivity effect of the investment (the marginal Efficiency of Investment) and (D) the resource allocation effect. The resource allocation effect (D) depends on parental preferences α_s. The total effect of the shock (A) will be less than the biological effect (B) in case parents respond with investments to counter the adversity of the early-life shock. High values α_s imply more weight to child outcomes and imply higher levels of parental investment to counter the early-life shock. Conversely, given a biological effect (B), a higher weight (α_s) for boys than for

[23] Almond and Mazumder (2013) give examples that show that compensation for an early-life shock is optimal in some settings and not in others. Moreover, the absence of behavioral responses could also be the result of a family being financially constrained. See Almond and Mazumder (2013) for more details.

girls implies more investments for boys than for girls and hence smaller reduced form effects of iodine deficiency for boys at later ages. This is in line with the findings of some empirical studies.

The salt iodization policy neutralizes geographically determined iodine deficiency rates. The policy has, therefore, two consequences. First, it reduces the negative cognition effects of iodine deficiency for both genders. In the presence of gender preferences, the reduction in female disadvantage in cognition due to the reform will be more substantial than for boys (holding the biological effect (B) constant). This is consistent with our finding of positive and significant cognition effects of the salt iodization policy for girls and small and insignificant effects for boys. Second, the salt ionization program will alter parental investment decisions. The need to invest in cognition is reduced or may even be absent, and parents may divert (part of the) investments into other dimensions of human capital (notably non-cognitive skills). For example, parents may set a particular target for schooling for their children. When after the salt iodization program targets for schooling are met, parents may shift some of their investment skill dimensions.

Although the literature on non-cognitive skills in developing countries is still very limited, there are several reasons to hypothesize that parents will invest in non-cognitive skills. Firstly, parents can easily observe emotions, mental and physical health. Secondly, parents might believe that non-cognitive skills could improve the child's welfare. Although there is little evidence about the returns to income from non-cognitive skills in a developing country, parents may be convinced that non-cognitive skills can benefit their child's welfare via other dimensions (e.g., the marriage market).

Although the literature on non-cognitive skills in developing countries is still limited, there are several reasons to hypothesize that the parental can and will respond to and invest in non-cognitive skills. Firstly, non-cognitive skills might be quite salient. Parents can easily observe their child's mental health and emotion, mainly targeted by the CES-D questionnaire. Secondly, parents might reasonably believe that non-cognitive skills could substantially affect children's welfare. Even there is very little evidence around the returns to non-cognitive skills in a developing country context in terms of income, parents can still reasonably conclude that non-cognitive skills can benefit their children in other dimensions (e.g., the marriage market).

Note that this does not imply that cognitive and non-cognitive skills are substitutes and that this behavior does not contradict skill complementarity in the sense of Heckman (2007).[24] Indeed, a prediction from the model (Yi, Heckman, Zhang, and Conti,

[24] It needs to be pointed out that here that dynamic complementarities, as discussed in the literature by Heckman (2007), only imposes restrictions on the functional form of the human capital production technology. Importantly, the dynamic process of human capital formation is jointly determined by parental preferences,

2015; Almond, Currie, and Duque, 2018) is that parents can compensate and reinforce initial shocks along different dimensions of human capital. Or, related, via "cross-productivity" (Cunha, Heckman, Lochner, and Masterov, 2006), changes in one dimension of human capital may also affect the accumulation of other dimensions.[25] Therefore, although the medical literature primarily documents neuro-developmental impairments as a consequence of in-utero exposure to iodine deficiency (Zimmermann, 2011), it cannot be ruled out that elimination of iodine deficiency also impacts non-cognitive dimensions of human capital (such as physical health, personality traits, risk attitude, etc.). The effect (sign and magnitude) of the salt iodization on non-cognitive outcomes is ultimately an empirical matter.

We use the baseline specification (3.1) to estimate the effect of the salt iodization policy on five different measures of non-cognitive skills. The first four measures are obtained from the 20 questions in the CES-D module of CFPS-2012. See Appendix C for details about the construction of the four measures. The fifth measure of non-cognitive skills is derived from a question in CPFS-2014: "How well are you getting along with others?". The results of the difference-in-differences regressions for the non-cognitive skill measures are displayed in Table 3.4. The table shows that the salt-iodization policy significantly improved retarded somatic activity, depressed affect, interpersonal problems, and social skills for boys. The corresponding coefficients for females are much smaller and insignificant. This finding is consistent with the idea that when the salt-iodization policy eliminates the cognitive disadvantage at birth, parental investments in boys that were initially geared towards cognitive skills are now diverted to other dimensions of human capital. Two alternative explanations are biological pathways and "self-productivity of skills" denoted by Cunha, Heckman, and Schennach (2010). To the best of our knowledge, there is no medical evidence that iodine deficiency in utero is related to the non-cognitive skills development. Therefore, the biological pathway is not likely to be the main explanation. And recall that we do not find significant impacts of the policy on boy's cognitive abilities, "self-productivity of skills" is also not likely to explain the positive effects on the non-cognitive skills of boys. However, it cannot be ruled out

budget constraints, and the production technology. So, even if there are dynamic complementarities in the technology of human capital formation such that "capabilities beget capabilities", parents can still respond endogenously by mitigating or reinforcing disadvantages early in life. Similar arguments have been made by Malamud, Pop-Eleches, and Urquiola (2016) and were confirmed by a series of empirical studies (Adhvaryu, Molina, Nyshadham, and Tamayo, 2015; Gunnsteinsson, Adhvaryu, Christian, Labrique, Sugimoto, Shamim, and West Jr, 2018; Duque, Rosales, and Sanchez, 2018; Aguilar and Vicarelli, 2018; Rossin-Slater and Wüst, 2018).

[25] Note, however, that in our simple model there is no role for "cross-productivity" and this is thus not captured in the decomposition (3.4). An example of decomposition which includes "cross-productivity" can be found in Grönqvist, Nilsson, Robling et al. (2018).

that primarily, the production technology of cognitive and non-cognitive skills and their interaction drives our findings (Almond, Currie, and Duque, 2018). We also used the total CES-D score as a measure for non-cognitive skills. We find a sizable and significant coefficient of -0.912 (with a p-value of 0.044).

Table 3.4: Iodine Exposure and Non-Cognitive Skills

	(1)	(2)	(3)	(4)	(5)
	Somatic Complaints	Depressed Affect	Positive Affect	Interpersonal Problems	Social Skills
Panel A: Males					
Post × Goiter	-0.354	-0.453	0.0612	-0.147	0.772
	[0.134]**	[0.162]**	[0.205]	[0.0765]*	[0.447]*
Mean of Dep. Var.	3.331	2.542	4.783	0.555	22.33
Observations	1216	1218	1218	1218	950
Panel B: Females					
Post × Goiter	0.104	-0.0147	0.111	-0.0689	-0.0150
	[0.195]	[0.233]	[0.151]	[0.0702]	[0.474]
Mean of Dep. Var.	3.539	3.259	4.898	0.657	22.46
Observations	1180	1180	1179	1180	915

Notes: Each coefficient is from a separate regression. All regressions control for fixed effects specific to birth province and birth year, birth order, family size, parents' education, region-specific linear trends and age. Mean-reversion controls include hospitals per capita in 1991, hospital beds per capita in 1991, and the sex ratio in Census1990 all interacted with cohort dummies. Standard errors clustered by province appear in square brackets. *, **, *** indicates significance at the 10%, 5% and 1% level respectively.
Data: CFPS-2012 and CFPS-2014.

3.6.2. GENDER ATTITUDES

Although differences in parental preferences can rationalize gender differences in the effects of the iodization policy, we cannot rule out the possibility that biological differences between men and women entirely drive all such gender differences. Therefore, in this section, we put the focus on within-gender analyses. We exploit plausibly exogenous variation on the resource allocation effect (D) generated by parental preferences (α_s in the utility function). Assuming that for a given gender, the biological effect (B) is independent of parental preferences (α_s), then the total reduced form effect (A) will not vary with parental preferences if there are no behavioral effects. On the other hand, reduced form estimates that vary with gender preferences hint at the relevance of behavioral responses of parents.

The assumption that gender preferences play no role in the biological effect requires more discussion. With the introduction of ultrasound techniques, parents could re-

spond prenatally, either with selective abortion or with increased antenatal investments when a male fetus is identified (see Bharadwaj and Lakdawala, 2013). In the latter case, we take the position that antenatal investments are included in parental investment decisions. As regards selective abortions, our empirical analyzes in the next section rely on within gender variation in parental preferences. This allows us only to need the much weaker assumption that for a given gender, the biological effect does not depend on α_s. We provide a test based on sex-ratios that supports this assumption later in section 3.6.3.

We do not directly observe gender preferences and rely on a plausibly exogenous proxy for parental gender preferences. We take the advantage that the CFPS-2014 includes a module with several questions on gender equity attitudes.[26] This module covers topics on gender roles within the household and in public life and asked respondents whether they agree with six statements phrased against gender equality and women empowerment. The response categories ranged from 1 "Strongly disagree" to 5 "Strongly agree". See Table C1 of the Appendix C for the statements and average responses for adult males and females. The gender attitude index at the two bottom rows are the mean and normalized mean of the individual responses for the six statements. A lower gender index means more gender-equitable views. Ideally, we would like to have gender attitudes measured 10 to 20 years before so that they coincide with the period during which investments were made. Unfortunately, this information is not available. Therefore, we approximate gender attitude measured in 2014, by which we have to assume that gender attitudes differences are relatively stable over time. For example, Attané (2012) found that the percentage of women(men) who agree with "Men are turned toward society, women devote themselves to their family" actually increased by only 4.4% (7.7%) from 50.4% (53.9%) from 2000 to 2010. The fact that percentage remains stable both in men and women reveals the deep-seated internalization of gender inequality.

Gender attitudes have drawn great interest in the recent literature (see Dahl, Kotsadam, and Rooth, 2017; Dhar, Jain, and Jayachandran, 2018; Dossi et al., 2019). For instance, Dossi, Figlio, Giuliano, and Sapienza (2019) explicitly link son preference to under-performance of girls in the U.S. We follow the design of Dossi, Figlio, Giuliano, and Sapienza (2019) to examine the association between gender attitudes and parental investment. The parental investment is measured by an unweighted average of the z-score of 5 variables: breastfeeding practices, whether the child went to kindergarten, how often the parent read to their child, how often the parent buys books for their child; and

[26] An alternative proxy could be the sex ratio (see Edlund et al., 2013, for instance). Selective abortion may be higher in regions where preferences for boys are stronger than in other regions. We do not use sex ratio as a proxy for gender preferences as it is not a measure of son preference *per se*, but rather the realization of the family's son preference combined with the preferences over the family size Jayachandran (2015).

how often they travel with their child. Unfortunately, who do not observe such parental investment score for the parents of our target sample of children born between 1991 and 2000. We do, however, observe the parental investment score for children younger than 6 in CPFS.

We regress the parental investment score on gender attitude, an interaction of gender attitude with gender, and a set of individual characteristics (Table C4 in Appendix C). We find gender attitudes to be strongly associated with parental investments in children, and boys' investments are higher in areas where preferences favoring boys are stronger. This finding remains robust when we exploit within province variation in the gender attitude variable by adding province fixed effects. This is important for the next subsection, where we exploit within province variation in gender attitudes to identify the effect of the iodization policy on cognitive and non-cognitive outcomes by gender.

3.6.3. GENDER ATTITUDES AND THE EFFECT OF THE SALT IODIZATION

Assuming that the biological effect does not depend on gender preferences α_s, the model in section 3.6.1 hypothesizes that systematic variation of the reduced form estimate (A) with gender preferences hints at an important role for parental gender preferences in human capital investments in children. Gender preferences can be incorporated by gender specific preference weights α_s and to test for this we specify the following triple-difference equation that we separately estimate per gender:

$$Y_{ijpt} = \beta_0 + \beta_1 Post_t \times Goiter_p \times GA_j + \beta_2 Post_t \times Goiter_p + \beta_3 Post_t \times GA_j + \beta_4 GA_j + X_{ipt}\rho + \delta_p + \gamma_t + \epsilon_{ipt} \tag{3.5}$$

Y_{ijpt} is the human capital outcome of child i, living in village/community j of province p at time t. $Post_t$, $Goiter_p$ and X_{ipt} are defined as before. The proxy for gender preferences GA_j is taken at the village/community level. The mean gender attitude for an individual i is calculated among adults born between 1951 and 1986 (about the same age as the parents of our target sample) for the village/community that child i resides in. One concern is that gender norms are effectively not randomly allocated across communities. Therefore, we also estimate the model with additional controls on communities' characteristics interacted with cohort dummies. We construct 15 communities' pre-policy characteristics from the village module of the CFPS-2010.[27] Estimates controlling for communities' characteristics doesn't alter the results (see Appendix C, Table C2). As in the main analyses, we condition on province fixed effects (δ_p). Therefore, our empirical specification absorbs differences in child human capital across provinces and solely relies on within province community/village level variation in gender preferences. By

[27] The information in village module was collected from a knowledgeable individual who has access to statistical materials in the village, such as the director or the accountant of the community committee.

conditioning on birth cohort fixed effects (γ_t), we aim to absorb all variation across age groups. We also include X_{ipt} to control for individual, family and provincial characteristics.

The main coefficient of interest is β_1, and, as we estimate the model separately by gender, the coefficient measures the within gender differential effect of the iodine fortification program across families with varying levels of son preferences. Mapping β_1 to the decomposition (3.4), the first term (B) on the right-hand side is the biological effect, which we assume to be homogeneous for a given gender and thus independent from parental gender preferences (α_s). As a consequence, β_1 will reflect the extent to which the behavioral response ($C \times D$) depends on α_s. We expect preferences to primarily have a role via D, the parental investment response to the salt-iodization program. The term C is the efficiency of the investment and depends on the health production technology. While C might be related to gender preferences α_s, it is expected that α_s plays a more direct role in D.[28] Note that the Gender Attitude variable is normalized with zero mean, and therefore the estimate of β_2 can be compared with the reform effect of section 3.5.

Table 3.5 reports the results by gender for cognitive outcomes. For females, high values of GA mean lower values of α_s in the theoretical model. The estimates of β_1 are reported in the first row of panel A (for males) and B (for females). The estimates of β_1 show large and significant effects (specifically for Math and Schooling) for females and (slightly) smaller and significant effects for males. Girls residing in villages/communities with strong preferences for boys benefit more from the universal salt iodization program than otherwise similar girls in communities/villages with less strong preferences for boys. The effects for boys is slightly smaller but with an opposite sign. Recall that high values of GA mean high values of α_s for boys in the theoretical model. Therefore, a negative estimate of β_1 suggests that boys from villages/communities with strong preferences for boys benefit benefit less from the universal salt iodization program. Estimates from both boys and girls suggest that the son preferences of parents drive at least part of the differential effects by gender of the effect of the salt iodization program on cognitive outcomes. Parental investments (D) in cognition in boys are partially crowded out by the universal salt-iodization program.

One concern for our triple-difference results in Table 3.5 is that the gender selection could also drive these results. With boy biased gender preferences, parents could selectively abort female fetuses. In this case, the marginal girl born in a village with strong boy preferences may differ from the marginal girl in a gender-neutral village. We link

[28] The The efficiency of the investment may vary with gender preferences if, for instance, parents do not monitor the investment's effect. It is expected that an effect of α_s on C (if present) is a second-order effect.

Table 3.5: The Impact of Iodine Exposure by Gender Attitudes

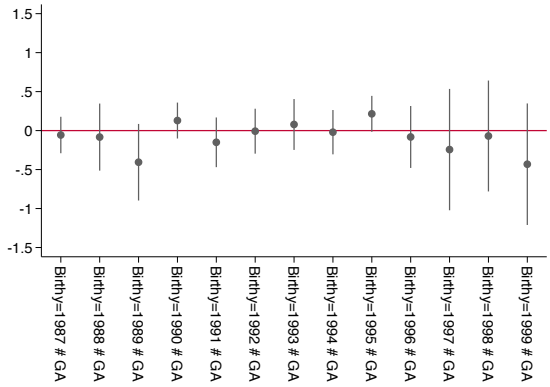
	(1)	(2)	(3)	(4)
	Math Test ln(scores)	Verbal Test ln(scores)	Primary School	Schooling ln(years)
Panel A: Males				
Post × Goiter × Gender Attitudes	-0.0617	-0.0748	-0.0613	0.00322
	[0.0335]*	[0.0535]	[0.0325]*	[0.0477]
Post × Gender Attitudes	-0.140	-0.143	-0.00764	-0.00815
	[0.137]	[0.123]	[0.0872]	[0.0886]
Post × Goiter	0.0533	0.0815	0.0311	0.0224
	[0.0498]	[0.0489]	[0.0247]	[0.0365]
Mean of Dep. Var.	2.517	3.083	0.497	2.107
Observations	3302	3302	3780	3644
Panel B: Females				
Post × Goiter × Gender Attitudes	0.176	0.0800	0.0333	0.0977
	[0.0558]***	[0.0409]*	[0.0249]	[0.0486]*
Post × Gender Attitudes	-0.0566	-0.0232	-0.0541	0.0620
	[0.175]	[0.129]	[0.144]	[0.261]
Post × Goiter	0.120	0.116	0.0342	0.0885
	[0.0391]***	[0.0536]**	[0.0103]***	[0.0220]***
Mean of Dep. Var.	2.543	3.171	0.529	2.122
Observations	3109	3109	3507	3383

Notes: Each coefficient is from a separate regression. All regressions control for fixed effects specific to birth province and birth year, birth order, family size, parents' education, region-specific linear trends, survey wave by age interactions and survey wave by cohort interactions. Mean-reversion controls include provincial average educational attainment in the 1990 Census interacted with the dummy for treated cohorts, hospitals per capita in 1991, hospital beds per capita in 1991, and the sex ratio in Census1990 all interacted with cohort dummies. Standard errors clustered by province appear in square brackets. *, **, *** indicates significance at the 10%, 5% and 1% level respectively.
Data: CFPS-2010 and CFPS-2014

the community/village level gender attitudes to sex ratio data to address this potential threat of endogenous sex selection. Ideally, we would like to construct sex ratio information at the community/village level. Unfortunately, this information is not available. Therefore, we rely on sex-ratio information at the city unit (the next level of spatial aggregation).[29] Like the event study design in section 3.5.4 we relate city level sex-ratios to gender attitudes by year, controlling for province fixed effects. The results of this event study are depicted in Figure 3.6. There is no association between gender attitude and sex ratios, and, importantly, the association doesn't change after introducing the salt iodization policy. Combining the evidence of this analysis with the event study analyses in section 3.5.4, we can conclude that sex selection is not a likely driver of our findings in Table 3.5.

[29]There are 128 city units, each consisting of on average four communities/villages.

Figure 3.6: Sex Ratio and Gender Attitudes



Notes: Sex ratios are aggregated from microdata of the 2000 Census. Following Edlund, Li, Yi, and Zhang (2013), we calculate the sex ratio at city level for each birth year between 1987 and 1999. In the event study regression, we control for province fixed effects and birth year fixed effect. Standard errors are clustered by city.

The salt-iodization program did not affect parental investment in cognitive skills for boys. As argued before (section 3.5.4), parents could divert their investments in cognitive skills for boys to other dimensions of human capital, notably non-cognitive skills such as physical health, mental health, social skills, non-cognitive skills that enhance labor market outcomes, etc. Indeed, the results in Table 3.4 showed that boys do benefit along some non-cognitive skill dimensions. To see whether the program effects on non-cognitive factors vary with gender preferences, we also estimated the triple-difference model for the set of non-cognitive variables available for our cohorts, born between 1991-2000. This includes the subscores of the CES-D test and self-reported social skills. The estimates in the third row of Panel A (Boys) and B (Girls) in the Table C3 are the effects of the implementation of the salt iodization program (β_2) and are in line with the findings in Table 3.4. However, none of the effect estimates of β_1 in the first row of Panel A and B are significant for our set of non-cognitive skill measures, suggesting that the effect of the salt-iodization program does not vary by gender attitude. An explanation for this can be found in the primary impact of the salt-iodization program. In utero exposure to iodine deficiency has effects on fetal brain development primarily, and therefore, the reform's first order effects are on cognition. The reform also affects the allocation of resources over the different dimensions of human capital. This second-order effect is reflected in the estimates of the impact of the policy (β_2). Therefore, to what extend the impact of the policy on non-cognitive skills for boys (or girls) varies within gender by

gender preferences (β_1) can be considered a third-order effect. Indeed, this effect might be small. It might be too demanding for the data and the measures of non-cognitive skills that we have at our disposal.

3.7. CONCLUSION

Currently, about 2 billion people suffer from iodine deficiency. The medical literature documents that iodine deficiency can lead to neurodevelopmental problems when fetuses are exposed in utero. This can lead to a reduction in children's cognitive skills and, consequently, adverse labor market outcomes later in life. This paper evaluates the effect of a nationally implemented salt iodization program on cognition of school-aged children in China. It differs from previous work as we explicitly focus on the role of gender preferences and how this may affect the effectiveness of large scale public programs. Gender preferences may also explain gender differences in the empirical literature on the long-run effects of adverse conditions early in life. Our difference-in-differences analyses find strong positive effects of the program for girls. A one standard deviation decrease (12%) in the pre-intervention children goiter rate is associated with math and vocabulary scores increasing by roughly 15%. We also see large increases in the educational attainment of females. Using the simple back of the envelope calculation, we infer that this translates to income increases of about 6,6%. Yet, we do not find any effects for boys.

These findings are robust against alternative specifications and falsification tests. These findings thus support the effectiveness of important, low costs, public health intervention. Indeed, compared to other interventions to raise education, the cost of salt iodization is extremely low. The costs associated with the intervention are about 0.05$ per person per year (WHO, 2005). This contrasts sharply with other interventions, such as class size reductions, costing over $5,000 (2010 dollars) per year per student (Chetty, Friedman, Hilger, Saez, Schanzenbach, and Yagan, 2011).

We proceed by further looking into these gender-specific findings and argue that gender preferences may play an important role. We do this in the context of a simple model along the lines of Yi, Heckman, Zhang, and Conti (2015); Almond, Currie, and Duque (2018) that explain how parental investments respond to (adverse) effects of shocks early in life. We show that gender preferences may lead to different investments in boys and girls. Following Yi, Heckman, Zhang, and Conti (2015); Almond, Currie, and Duque (2018) we use that total effect of an early life shock can be decomposed into a biological effect and a behavioral effect. The behavioral effect includes the parental investment decision, which mitigates or reinforces the biological effect. This behavioral

("resource allocation effect") depends on child preferences that may differ by gender. This simple model can explain the policy effect on girls and the absence of an effect on boys. Prior to the salt iodization policy, parents may have countered initial adverse shocks for boys and less so for girls. Therefore, when boy preferences are important, girls may benefit more from a nationally implemented programs. This also suggests that the salt-iodization program crowds out private parental investments in cognition. Another consequence of this may be that parents divert their investments into other skill dimensions, notably for boys. Indeed, we find program effects on non-cognitive skills for boys, but not for girls.

We test this hypothesis by explicitly accounting for gender preferences and proxy gender preferences with an index for gender attitudes. We estimate a triple-difference model per gender. Hence, the model identifies the effect of gender preferences from within gender variation in community/village gender attitudes. In line with the model and expectations, girls' program effects on cognition are stronger in communities/villages where preferences favoring boys are stronger. We do not find such effects for boys. Nor do we find such triple-difference effects for non-cognitive skills, both for boys and girls. Cultural or other contextual factors that feed attitudes favoring boys are not restricted to China but also hold for other countries in South-East Asia, the Middle East, and North Africa. The idea that gender preferences play a role in the western world can not be excluded either. Our findings, therefore, do not only speak to the external validity of the current study but also suggest that later life gender differences in labor market outcomes observed in many countries may be rooted early in life. The findings also indicate that generally large-scale programs may reduce gender inequalities and contribute to gender convergence.

Preferences favoring boys may lead to unequal investments in girls and justify policies that aim to reduce the consequences of such preferences. Also, improved economic circumstances may reduce such gender preferences, but as earlier work has shown (Almond and Edlund, 2008; Abrevaya, 2009) gender biases persist even with improved economic circumstances. This then calls for policies such as increasing mandatory schooling for both genders and other public programs that affect all, such as the public health campaign in this paper. Parental preferences are important, but this does not exclude the relevance of other factors such as financial constraints and the production technology of human capital formations. Indeed, for a significant part, these general investments can only improve outcomes as long as the investments translate into human capital; in terms of the model, the efficiency of investment (i.e., the term C in the decomposition (3.4)). With low values in the efficiency of investment, the impact of any investment (public or

private) remains modest. The key to understanding the factors that drive the efficiency of the investment effect lies in understanding the human capital production technology. Indeed, this is one of the topics that require further development (see, e.g. Cunha, Heckman, and Schennach, 2010; Agostinelli and Wiswall, 2016; Attanasio, Meghir, and Nix, 2018). Unfortunately, with the data at hand, we can not competently tackle this problem and therefore leave this to future research.

A. ADDITIONAL INFORMATION

In the analyses, we use a series of province-level pre-policy characteristics. We list all these variables here and describe how we construct each variable.

- Hospital per capita: Number of hospital per 1000 (Statistics yearbook 1991).
- Hospital beds per capita: Number of hospital beds per 1000 (Statistics yearbook 1991).
- School per capital: Number of primary schools per 1000 (Statistics yearbook 1991).
- Sex ratio: Sex ratio of the cohort 1991 calculated from Census 2000.
- Rural income: Average rural household income calculated from NFPS 1991.
- Poverty: Poverty rates in 1993 are obtained from Woo, Li, Yue, Wu, and Xu (2004).

Table A1: Regional Classification of Provinces

Region	Provinces
North China	Beijing, Tianjin, Hebei, Shanxi and Inner Mongolia
Northeast China	Liaoning, Jilin and Heilongjiang
East China	Shanghai, Jiangsu, Zhejiang, Anhui, Fujian, Jiangxi and Shandong
South Central China	Henan, Hubei, Hunan, Guangdong, Guangxi and Hainan
Southwest China	Chongqing, Sichuan, Guizhou, Yunnan and Tibet
Northwest China	Shaanxi, Gansu, Qinghai, Ningxia and Xinjiang

Notes: Author's tabulations

Table A2: Summary Statistics

	High Goiter Provinces		Low Goiter Provinces	
	Females	Males	Females	Males
Outcomes				
Educational Attainment	2.02	1.94	2.04	1.95
	[0.79]	[0.79]	[0.83]	[0.79]
Illiterate	0.28	0.33	0.30	0.32
Primary School	0.46	0.42	0.40	0.44
Middle School	0.24	0.24	0.28	0.22
High School or above	0.028	0.014	0.029	0.024
Schooling	7.35	7.08	7.49	7.25
	[2.52]	[2.45]	[2.51]	[2.42]
Math Test Scores	12.2	11.9	12.5	12.2
	[5.23]	[5.25]	[5.13]	[4.79]
Verbal Test Scores	23.8	22.3	24.2	22.5
	[7.32]	[7.64]	[7.01]	[7.29]
Demographics				
Age	14.5	14.4	14.3	14.4
	[2.52]	[2.50]	[2.48]	[2.48]
Father's Educational Attainment	2.19	2.20	2.46	2.48
	[0.94]	[0.95]	[0.88]	[0.88]
Mother's Educational Attainment	1.68	1.69	2.16	2.13
	[0.85]	[0.83]	[0.92]	[0.88]
Birth Order	1.56	1.69	1.63	1.72
	[0.73]	[0.86]	[0.96]	[0.92]
Family Size	5.13	4.80	5.10	4.78
	[1.57]	[1.45]	[1.66]	[1.54]
Number of observations	956	1028	997	1007

Notes: Author's tabulations of CFPS-2010 and CFPS-2014. Sample consists individuals born in rural area between July 1995 and June 2000. We label a province as high/low goiter if its goiter rate is above/below 17% (median).

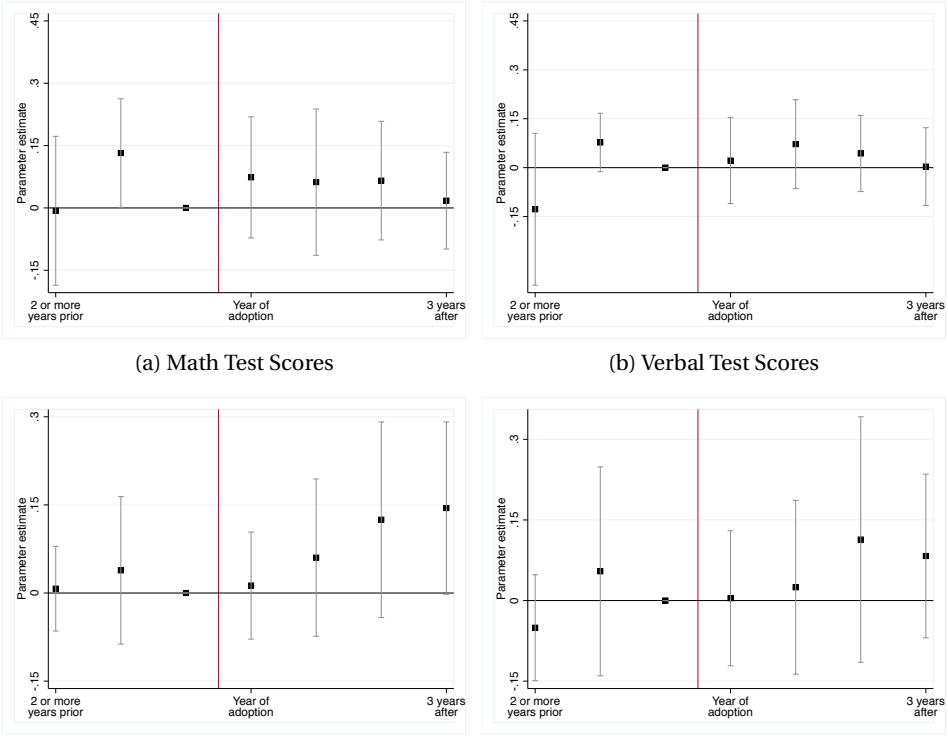
B. ADDITIONAL RESULTS: IODINE AND LONG-RUN OUTCOMES

Table B1: Robustness Checks (Male)

	(1) Math Test ln(scores)	(2) Verbal Test ln(scores)	(3) Primary School	(4) Schooling ln(years)
Panel A: raw test scores				
Post × Goiter	0.108	0.405	0.0202	0.0878
	[0.333]	[0.473]	[0.0233]	[0.169]
Mean of Dep. Var.	13.59	23.72	0.499	8.569
Observations	3310	3310	3803	3654
Panel B: additional controls				
Post × Goiter	-0.0588	-0.0580	-0.00554	-0.0143
	[0.0507]	[0.0538]	[0.0297]	[0.0686]
Mean of Dep. Var.	2.517	3.083	0.499	2.108
Observations	3310	3310	3803	3654
Panel C: drop partial exposed group				
Post × Goiter	0.0296	0.0575	0.0260	0.0216
	[0.0517]	[0.0524]	[0.0230]	[0.0403]
Mean of Dep. Var.	2.514	3.081	0.497	2.106
Observations	3235	3235	3706	3560
Panel D: only using baseline wave 2010				
Post × Goiter	0.0551	0.0809	-0.00890	0.0111
	[0.0567]	[0.0666]	[0.0290]	[0.0369]
Mean of Dep. Var.	2.486	3.046	0.344	1.978
Observations	1992	1992	2028	2028
Panel E: small sample window				
Post × Goiter	0.0633	0.0668	0.0223	0.0344
	[0.0584]	[0.0660]	[0.0213]	[0.0465]
Mean of Dep. Var.	2.549	3.108	0.516	2.129
Observations	2643	2643	3047	2920

Notes: Each coefficient is from a separate regression. All regressions except Panel B use the same controls as the baseline model in Table 3.2. In Panel B, we control for birth-region and birth-year specific interaction instead, and we additionally control for schools per capital in 1991, poverty rates in 1993, and average household income in 1991 all interacted with cohort dummies. Standard errors clustered by province appear in square brackets. *, **, *** indicates significance at the 10%, 5% and 1% level respectively.
Data: CFPS-2010 and CFPS-2014

Figure B1: Event Study (Male)



(a) Math Test Scores

(b) Verbal Test Scores

(c) Primary School

(d) Schooling

Notes: The sample includes all male respondents from two waves of the survey (CFPS-2010 and CFPS-2014). Each point reflects the coefficient estimated on an interaction term between the birth year (compared to 1995) and the pre-intervention (base) level of the goiter rate in the birth-province. Capped spikes represent 95 percent confidence intervals. All models condition upon birth province and birth year fixed effects and the full set of controls used in our main analysis.

C. ADDITIONAL RESULTS: MORE ON GENDER

IODINE EXPOSURE AND NON-COGNITIVE SKILLS

CES-D includes 20 questions, which were aggregate to four categories: somatic complaints (Q1, Q2, Q5, Q7, Q11, Q13, Q20); depressed affect (Q3, Q6, Q9, Q10, Q14, Q17, Q18); positive affect (Q4, Q8, Q12, Q16)[30]; interpersonal problems (Q15, Q19).

Questions: Below is a list of the ways you might have felt or behaved. Please tell me how often you have felt this way during the past week with a score: 1 for rarely or none of the time (less than 1 day); 2 for some or a little of the time (1-2 days); 3 for occasionally or a moderate amount of time (3-4 days); 4 for most or all of the time (5-7 days).

1. I was bothered by things that usually don't bother me.
2. I did not feel like eating; my appetite was poor.
3. I felt that I could not shake off the blues even with help from my family or friends.
4. I felt I was just as good as other people.
5. I had trouble keeping my mind on what I was doing.
6. I felt depressed.
7. I felt that everything I did was an effort
8. I felt hopeful about the future.
9. I thought my life had been a failure.
10. I felt fearful.
11. My sleep was restless.
12. I was happy.
13. I talked less than usual.
14. I felt lonely.
15. People were unfriendly.
16. I enjoyed life.

[30] We reverse the score of each question in the positive affect according to the literature (i.e., 1 means most or all of the time (5-7 days) and 4 means rarely or none of the time (less than 1 day); 2 for occasionally or a moderate amount of time (3-4 days); 3 for some or a little of the time (1-2 days)).

17. I had crying spells.

18. I felt sad.

19. I felt that people dislike me.

20. I could not get "going".

Table C1: Summary Statistics of Gender Attitudes

	Males	Females
Son should live together with his parents.	3.49	3.47
	[1.34]	[1.38]
Every family should at least have a son.	3.46	3.36
	[1.48]	[1.55]
The husband takes care of the business, and the wife takes care of the family.	4.08	4.07
	[1.10]	[1.15]
Woman's marriage is more important than her career.	3.50	3.75
	[1.33]	[1.27]
Every woman should have a child.	4.08	4.33
	[1.17]	[1.02]
Disagree: The husband should do half of the housework.	1.95	1.84
	[1.10]	[1.05]
Gender index	3.43	3.47
	[0.70]	[0.71]
Normalized gender index	-0.031	0.030
	[0.99]	[1.01]
Number of observations	10070	10455

Notes: Sample includes all individuals born between 1951 and 1986 in CFPS-2014. Surveyed respondents were asked if they agree with these six statements. The respondents report how much they agreed with a certain statement on a scale of 1-5, with 1 being *Strongly agree* and 5 being *Strongly disagree*. Gender index is the average of the 6 indicators. Normalized gender index is calculated by subtracting the mean and dividing by the standard deviation.

ADDITIONAL RESULTS ON GENDER ATTITUDES

To show that parental attitudes reflect parental preferences is related to parental investment in children, we exploit the fact that CFPS does include some information about parental investments behavior for a different sample of children younger than 6 years of age. For this sample we construct a "parental investments index" by calculating an unweighted average of the z-score of 5 variables: breastfeeding practices; whether the child went to kindergarten; how often the parent read to their child; how often the parent buys books for their child; and how often they travel with their child.[31] Breastfeeding

[31] Iodine could also be obtained from food such as vegetables, eggs, fish and meat. The composition of the diet as well as the quantity of food may therefore influence the iodine intake of pregnant women and infants.

Table C2: The Impact of Iodine Exposure by Gender Attitudes

	(1)	(2)	(3)	(4)
	Math Test ln(scores)	Verbal Test ln(scores)	Primary School	Schooling ln(years)
Panel A: Males				
Post × Goiter × Gender Attitudes	-0.0757	-0.102	-0.0518	-0.0284
	[0.0430]*	[0.0607]	[0.0423]	[0.0447]
Post × Gender Attitudes	-0.0931	-0.0180	0.105	0.105
	[0.153]	[0.110]	[0.0928]	[0.138]
Post × Goiter	0.0310	0.0689	0.0235	0.0112
	[0.0475]	[0.0445]	[0.0227]	[0.0360]
Mean of Dep. Var.	2.517	3.082	0.497	2.107
Observations	3257	3257	3731	3598
Panel B: Females				
Post × Goiter × Gender Attitudes	0.136	0.0209	0.0517	0.0750
	[0.0671]*	[0.0599]	[0.0301]*	[0.0433]*
Post × Gender Attitudes	0.0124	0.0644	-0.0741	0.171
	[0.217]	[0.157]	[0.137]	[0.284]
Post × Goiter	0.110	0.125	0.0307	0.0857
	[0.0339]***	[0.0499]**	[0.0124]**	[0.0208]***
Mean of Dep. Var.	2.541	3.171	0.528	2.122
Observations	3062	3062	3457	3334

Notes: Each coefficient is from a separate regression. All regressions control for fixed effects specific to birth province and birth year, birth order, family size, parents' education, region-specific linear trends, survey wave by age interactions and survey wave by cohort interactions. Mean-reversion controls include provincial average educational attainment in the 1990 Census interacted with the dummy for treated cohorts, hospitals per capita in 1991, hospital beds per capita in 1991, and the sex ratio in Census1990 all interacted with cohort dummies. Additionally, communities controls include cohort dummies interacted with 15 communities' pre-policy (1991) characteristics: access to electricity, radio, satellite TV, post services, phone, mobile phone, highway, rail, tap water, gas, local factory, hospital, local election, urbanization, and land property rights. Standard errors clustered by province appear in square brackets. *, **, *** indicates significance at the 10%, 5% and 1% level respectively.
Data: CFPS-2010 and CFPS-2014

is a dichotomous variable whether the duration of breastfeeding was longer than three months. Kindergarten is an dichotomous variable indicating whether the child has attended kindergarten. The last three variables are categorical variables for the frequency of the event ranging from 0 ("never") to 5 ("every day"). We take higher values of the score to be associated with more investments in the child. It is not obvious that this also holds for the kindergarten indicator. Recent papers on the effect of day care (children aged 0-2) find mixed results. Positive effects are found for Norway (Drange and Havnes (2019) shown signicant gains in language and mathematics at age 6–7 of childcare enrollment

The eating of healthy and nutritionally rich food could be viewed as an antenatal investment and may differ with knowledge of the gender of the unborn child. Unfortunately we do not have information on food consumption of pregnant women and infants.

Table C3: Iodine Exposure and Non-Cognitive Skills

	(1)	(2)	(3)	(4)	(5)
	Somatic Complaints	Depressed Affect	Positive Affect	Interpersonal Problems	Social Skills
Panel A: Males					
Post × Goiter × Gender Attitudes	0.143	-0.288	-0.680	-0.139	0.737
	[0.243]	[0.331]	[0.292]**	[0.119]	[0.649]
Post × Gender Attitudes	-0.740	-0.655	0.287	-0.0193	0.320
	[1.267]	[0.971]	[1.534]	[0.452]	[2.525]
Post × Goiter	-0.487	-0.438	0.121	-0.0852	0.710
	[0.197]**	[0.166]**	[0.168]	[0.0496]*	[0.346]*
Mean of Dep. Var.	3.340	2.587	4.808	0.576	22.29
Observations	1554	1556	1556	1555	1064
Panel B: Females					
Post × Goiter × Gender Attitudes	-0.559	-0.289	-0.129	-0.245	0.495
	[0.418]	[0.365]	[0.325]	[0.182]	[0.835]
Post × Gender Attitudes	-0.415	0.519	0.107	0.544	-0.00960
	[1.305]	[1.105]	[0.902]	[0.407]	[2.011]
Post × Goiter	-0.0797	-0.114	0.0518	-0.00230	-0.180
	[0.217]	[0.299]	[0.172]	[0.0726]	[0.350]
Mean of Dep. Var.	3.515	3.259	4.872	0.653	22.43
Observations	1436	1436	1435	1436	1017

Notes: Each coefficient is from a separate regression. All regressions control for fixed effects specific to birth province and birth year, birth order, family size, parents' education, region-specific linear trends and age. Mean-reversion controls include hospitals per capita in 1991, hospital beds per capita in 1991, and the sex ratio in Census1990 all interacted with cohort dummies. Standard errors clustered by province appear in square brackets. *, **, *** indicates significance at the 10%, 5% and 1% level respectively.
Data: CFPS-2012 and CFPS-2014.

using childcare assignment lotteries.) and Germany (Felfe and Lalive (2018) shown early child care has a strong positive effects on children's motor and socio-emotional skills.) while the recent paper by Ichino, Fort, and Zanella (2019) find negative effects, in particular for girls.[32] Note that kindergarten usually concerns older infants (age 4 and older). Also, for our sample of rural families interaction with other infants may be more beneficial and can be viewed as an investment.

For this sample of parents who have children younger than 6, we run a regression that relates the parental investment index to our proxy for gender preferences:

$$I_{ij} = \beta_0 + \beta_1 Male_i \times GA_j + \beta_2 Male_i + \beta_3 GA_j + X_i\rho + \epsilon_i, \tag{3.6}$$

[32] They examine the effects of extended day care of children aged 0-2 on cognitive and non-cognitive skills. The idea is that daycare implies fewer one-to-one interactions with adults as inputs in the technology of skill formation and therefore be more harmful for infants. Girls at age 0–2 are relatively more capable of making good use of stimuli that improve their skills. Extended exposure to daycare may therefore be particularly harmful for girls.

where I_{ij} is parental investments index for child i from village/community j. $Male_i$ is dummy variable indicating the gender of the child, and GA_j is the gender attitude at the village/community j.[33] Similar to Dossi, Figlio, Giuliano, and Sapienza (2019), we also control for a series of variables X_i such as parental education, birth order, family size and province fixed effects. Prime interest is in the parameter β_1, the association between gender preferences and parental investment in young children.

The results of the regression are reported in Table 3.5. In first three columns, we gradually add family background controls to the regression. In the fourth column, we also control for province fixed effects. In this specification the estimates solely rely on within province variation in gender attitudes. The estimates of β_1 are reported in the first row of Table 3.5 and show that gender attitudes are strongly associated with parental investments in children and that investments in girls are lower in areas where preferences favoring boys are stronger. Adding controls does not affect the estimate of β_1. The small change in β_1 when we add province fixed effects (column 4) also underlines that there is substantial within province variation in the gender attitude variable. This is important for the section 3.6.3 where we exploit this variation to identify the role of preferences in the effect of the iodization policy on cognitive and non-cognitive outcomes.

Table C4: Gender Attitudes and Parental Investment

	(1)	(2)	(3)	(4)
	Parental Investment Index			
Gender Attitudes × Male	0.163	0.189	0.186	0.165
	[0.0834]*	[0.0816]**	[0.0824]**	[0.0846]*
Male	-0.0665	-0.0495	-0.0489	-0.0381
	[0.0329]**	[0.0325]	[0.0321]	[0.0330]
Gender Attitudes	-0.315	-0.248	-0.221	-0.170
	[0.0771]***	[0.0711]***	[0.0728]***	[0.0791]**
Parental Education	No	Yes	Yes	Yes
Birth Order & Family Size	No	No	Yes	Yes
Province Fixed Effects	No	No	No	Yes
Observations	1,642	1,642	1,642	1,642

Notes: Each column is from a separate regression. Standard errors clustered by village/community appear in square brackets. *, **, *** indicates significance at the 10%, 5% and 1% level respectively.
Data: CFPS-2010, CFPS-2012 and CFPS-2014

[33] With this regression we do not aim to make causal statements. Nevertheless, in the calculation of the village/community level average we left out the individual response to minimize confounding.

4

MONITOR [1]

4.1. INTRODUCTION

Lacking enforcement of existing regulations is widely regarded as a serious impediment to welfare in developing countries. This is particularly the case for environmental policy. Despite a substantial increase in both the number and the strictness of environmental regulations across the globe during the past 50 years, enforcement of these laws has been limited (UN, 2019). According to the UN (2019), this lack of enforcement is one of the greatest obstacles to combat climate change and pollution. Since most regulations are set at the national (or international) level, and the implementation of policies is typically left to local officials – this leads to a classical principal-agent problem. Holding officials accountable in this setting is particularly challenging for both central governments and citizens since information on local environmental outcomes is often missing or of poor quality. One explanation for this is that information is often provided by local officials themselves, who have incentives to misreport (Banerjee, Duflo, and Glennerster, 2008; Fisman and Wang, 2017; Acemoglu, Fergusson, Robinson, Romero, and Vargas, 2018).

In this paper, we investigate whether better monitoring of the environment can strengthen local officials' accountability and improve environmental outcomes. We study this issue in the context of air pollution monitoring in China – a setting that we argue is particularly important. Combating air pollution is a global policy priority motivated by well-documented health and productivity consequences of exposure to pollution (Neidell

[1] This chapter is based on Axbard (2020).

and Currie, 2005; Greenstone and Hanna, 2014; Ebenstein, Fan, Greenstone, He, and Zhou, 2017; Jia, 2017; Barwick, Li, Rao, and Zahur, 2018). Despite international efforts in recent years to improve air quality, 91% of the world's population in 2018 still lived in areas where air pollution exceeded WHO guidelines. A large part of this population is living in emerging economies, including China, where pollution levels have exceeded the highest levels ever recorded in rich countries.

To study this issue, we exploit the introduction of more than 550 air pollution monitors in 177 Chinese prefecture-level cities in 2015.[2] Each monitor provides high-quality information to the central government on a range of different air pollutants close to the monitor. Hence, the larger the number of monitors installed in a city, the better able it is for the central government to track overall air quality in a city. Local officials, in turn, have incentives to decrease pollution close to the monitors since readings from the monitors feed into their performance evaluations, and these evaluations are based on achieving predetermined pollution reduction targets.

We start by estimating the causal impact of the number of monitors on actual air pollution. In general, there are two main empirical challenges to overcome. The first challenge is related to data availability: before the introduction of monitors, there is no reliable information on local pollution. To overcome this issue, we use satellite data that covers the period before and after the introduction of the monitors to capture changes in pollution.[3] The second challenge is related to identification: the number of monitors that a city chooses to install is potentially endogenous. For example, cities with more ambitious environmental goals might choose to install more monitors to get better information about the distribution of pollution. To overcome this challenge, we exploit a government program that assigns monitors to cities based on their population and geographical size. Using this information, we estimate a difference-in-differences model as well as instrumental variables and a regression discontinuity design using the assignment criteria set up by the central government. All three empirical strategies produce consistent results, and our preferred DiD estimate shows that one additional monitor reduces pollution by about 3%. This is a sizeable effect given that the median number of monitors in our sample is 3.

To bring clarity to how this reduction in pollution came about, we construct a novel data set of all air pollution enforcement activities by city governments from 2010 to 2017. We then use similar estimation strategies to those above to document that an additional monitor leads to a 20% increase in the number of regulatory enforcement activities. To

[2] The prefecture-level city is an administrative division ranking below a province and above a county. Figure B1 shows a picture of the type of monitors that we study.

[3] We provide validation of the satellite data using ground station measures in subsequent periods.

further validate that the monitors indeed drive this increase, we geo-reference enforcement activities and show that the increase in enforcement is driven by firms located within 20km from a monitor, whereas firms beyond 20km face no change in enforcement.

We explore two potential mechanisms for why monitors increase enforcement and reduce pollution: political promotion incentives and citizen pressure. The first mechanism is plausible in this setting since local mayors' performance is evaluated based on their achieved reduction of pollution. To assess this first mechanism, we follow Xi, Yao, and Zhang (2018) and exploit discontinuities in promotion incentives caused by the age of local mayors at the time of the National Congress (mayors can be promoted if they are below the age of 57, but not otherwise). The second mechanism is also a potential driving force of what we find since recordings of pollution data are made publicly available on the website of the Ministry of Environmental Protection. Since more and more empirical studies (Chen, Pan, and Xu, 2016) have found that authoritarian regimes are also responsive to citizen pressure, it is plausible that citizens better informed about pollution will put pressure on the local government to act. This mechanism is evaluated by using data on online searches for pollution-related keywords. We find strong evidence that the results of both enforcement and reduction in pollution are driven by mayors with incentives to be promoted, while we find limited evidence that monitoring increases citizens' awareness of pollution.

Finally, to understand how the quality of the information provided by the monitors' matters, we study the reassignment of control of the monitors from the local government to external third parties. This reassignment was conducted after it was discovered that several local governments tried to manipulate the data from the monitors. By exploiting information from the monitors and satellites, we show that the monitor recordings are more strongly correlated with the satellite data when they are under the control of a third party - suggesting a reduction in manipulation. Following this logic, we further document that when monitors are under the control of an independent third party, the effect of an additional monitoring on enforcement and reductions in pollution are substantial.

The paper relates to three strands of literature. First, the paper relates to a growing empirical literature studying pollution reduction in developing countries. Several papers in this literature have investigated the effect of changes in environmental policies (Greenstone and Hanna, 2014; Tanaka, 2015; Kahn, Li, and Zhao, 2015; Bombardini and Li, 2016; Ebenstein, Fan, Greenstone, He, and Zhou, 2017). These studies commonly use data from pollution monitors, and hence a key contribution of this paper is establishing

the impact of monitors themselves on pollution levels. If this effect is not taken into account when evaluating environmental policies, conclusions might be misleading.[4] Besides, we contribute to this line of work by directly investigating government enforcement activities. While prior work has recognized that enforcement is a major challenge (see, e.g., discussion in Greenstone and Hanna, 2014), we document how enforcement can be strengthened. Finally, close to our study is also the work by Duflo, Greenstone, Pande, and Ryan (2013, 2018), studies how changes to the incentive structure for third party auditors of polluting firms in India affect emissions. While the third-party auditors studied in these papers have a similar role to the private firms to which monitoring is assigned in this paper, the main difference is that we are concerned with how monitoring affects pollution reduction and data manipulation within the government rather than the firm.

Second, we contribute to studies showing that the provision of information can improve accountability and government performance (see review by Kosack and Fung, 2014). While most of this literature focuses on providing information to citizens about the performance of their politicians (e.g. Besley and Burgess, 2002; Snyder and Strömberg, 2010), this study is most closely related to the part of the literature that improves the flow of information from local to central government. This includes the work by Olken (2007) on the effect of top-down monitoring on corruption in village road projects in Indonesia as well as the studies by Reinikka and Svensson (2005, 2011) showing that providing information about schooling funds that local communities were supposed to achieve can reduce leakage of funds and improve test scores. Besides, it relates to studies of interventions aimed at improving the monitoring of local officials (e.g. Duflo, Hanna, and Ryan, 2012; Dhaliwal and Hanna, 2017). Our main contribution to this literature is acknowledging the role that local governments have in everyday information provision and how that affects top-down accountability. Besides, we study an autocratic setting where citizen accountability arguably plays less of an important role.

Third, it relates to literature investigating the potentially distorting effect of high-powered incentives. Studies have documented that such incentives can lead to manipulation of data (Banerjee, Duflo, and Glennerster, 2008; Fisman and Wang, 2017; Acemoglu, Fergusson, Robinson, Romero, and Vargas, 2018), including reporting on pollution (Andrews, 2008; Chen, Jin, Kumar, and Shi, 2013; Ghanem and Zhang, 2014; Oliva, 2015). We contribute to this literature by going beyond establishing that manipulation exists to study if and how it can be mitigated and whether such mitigation strategies

[4] This would, for example, be the case if pollution levels around the monitors are not representative of the overall pollution in a country.

affect actual policy performance (pollution reduction in this setting). This is made possible since we can observe not only the potentially manipulated data (reported from the monitors) but also the true outcome (observed via satellite).

The paper is structured as follows. Section 4.2 describes the context as well as the roll-out of the pollution monitoring program investigated. After that, the data used in this study is described (Section 4.3). The first analysis, which aims at estimating the causal effect of pollution monitoring on actual pollution, is presented in Section 4.4. This section discusses both the empirical strategy as well as the results. The analysis of the enforcement of environmental regulations is developed in 4.5, and mechanisms are discussed in Section 4.6. Finally, Section 4.7 offers concluding remarks.

4.2. INSTITUTIONAL CONTEXT

This section provides background information and describes why the national monitoring program studied in this paper was introduced. In the first sub-section (4.2.1), we describe the environmental policies in place in China during this period and discuss the role of local leaders in achieving these. After that, the implementation of the program is described in detail in Section 4.2.2.

4.2.1. ENVIRONMENTAL POLICIES IN CHINA

The concentration of air pollutants in China is among the highest in the world and a problem with serious health consequences. Average $PM_{2.5}$ (particulate matter with a diameter of 2.5 μm or less) concentrations in 2013 were 91 $\mu g/m^3$, which is nine times the amount that the World Health Organization considers safe. Estimates by Greenstone and Schwarz (2018) suggest that if such levels of pollution are sustained, it will result in a 6.5 years decline in life expectancy for the average resident. While the government's priority during the past decades has largely been to stimulate economic growth, attention has lately shifted towards environmental policies. Air pollution monitors have been installed through a nationwide program to track the impact of such policies.

Starting in 2013, the National Air Quality Action Plan was set up to improve air quality by the end of 2017. Specific targets of the plan were gradually laid out. The first broad goal of the plan was to lower urban concentrations of PM_{10} (particulate matter with a diameter of 10 μm or less) by 10 percent relative to 2012 levels. The second ambition was to reduce concentrations of $PM_{2.5}$ in the Beijing-Tianjin-Hebei provinces, the Yangtze River Delta and the Pearl River Delta by 25, 20, and 15 percent, respectively. After the plan was set up, the government took concrete steps to follow through on the broad goals outlined in the Action Plan. One particular action taken by the central government was

to set up pollution reduction targets for all provinces outside the three areas above. In January 2014, the Ministry of Environmental Protection (MEP) entered into "contracts" with all 31 provinces and set up a three-year air quality plan to decrease PM concentrations in the whole country.[5] In each "contract" an air quality target for 2017 was set – resulting in different percentage reduction targets of $PM_{2.5}/PM_{10}$ for each province.[6]

The implementation of these targets is carried out by local government units while the national government sets targets. To ensure efficient implementation of the targets, local officials are incentivized through performance-based promotions. Promotions are the key instrument used in China to ensure that local officials carry out policies in line with the goals set up by the central government (see Zheng and Kahn, 2013, 2017, for further discussion). For a long time, the central government focused on economic performance and emphasized economic growth as the key evaluation criteria for local officials' promotion opportunities (Chen, Li, and Lu, 2018). In 2011, the 12^{th} Five-Year Plan set the goal of reducing energy consumption. Shortly after that, the central government included environmental protection as one of the criteria when evaluating local governors' performance. Since then, the performance in terms of environmental protection is considered in promotion decisions of local mayors (Zheng and Kahn, 2013). As discussed in the previous section, the National Air Quality Action Plan (2013) imposed explicit targets to improve air quality by the end of 2017 for each province. However, it is hard to verify whether the targets set by the central government have been obtained without credible information about air pollution. Facing this challenge, the central government decided to expand and upgrade the national air quality monitoring system, the focus of this study.

4.2.2. NATIONAL MONITORING SYSTEM

The first air pollution monitoring in China began in the 1980s and was subsequently expanded during the next decade. In the 1990s, the number of cities covered had increased to 46. However, the data from the monitoring system during this period were never made public. It was not until 2000 that daily reports of air pollution data for 42 cities were published for the first time on the website of the Ministry of Environmental Protection (MEP). By 2012, the year before the program we study was launched, 113 cities, all provincial capitals, and some large industrial cities were required to measure three pollutants (i.e., PM_{10}, SO_2 and CO) and the measures of each pollutant were published online. Based on these measures, an Air Pollution Index (API) was constructed and

[5] The targets remained unchanged for provinces in Beijing-Tianjin-Hebei, the Yangtze River Delta and the Pearl River Delta.

[6] For the list of targets by province, see Table A1 in Appendix A.

disclosed to the public. The central government aimed to use this information disclosure to create an incentive for local governments to engage in air pollution reduction. However, the usefulness and reliability of this data has been questioned. First and foremost, it does not capture smaller particulate matter ($PM_{2.5}$), which is the most detrimental for health. Second, since monitors were only installed in large cities, a substantial part of the Chinese population was not covered by the monitors. Third, a discrepancy between the MEP data and measurements reported by the U.S. embassy in Beijing (which started measuring and publishing data on $PM_{2.5}$ in 2008) further questions whether the data captured actual pollution levels. This concern has later been validated in academic work (see, e.g. Ghanem and Zhang, 2014; Greenstone et al., 2019).

Upgrade and Expansion between 2013 - 2015 To address these issues, the government introduced a new monitoring system in its 2013 National Air Quality Action Plan. Following revised air pollution standards implemented in 2012, which included three additional pollutants ($PM_{2.5}$, O_3, and NO), existing monitors were updated. At the same time, the monitor network was also expanded to cover all prefecture-level cities that previously had no systematic air pollution monitoring in place. It is this expansion that we focus on in this paper. One of the key features of these new monitors is the ability to accurately record information on $PM_{2.5}$, widely regarded as the key measure of ambient air pollution. The second key feature is that all monitor stations report all six pollutants to the central government in real-time (Greenstone, Guojun, Ruixue, and Tong, 2019). Additionally, the hourly pollution data is automatically published online by the central government.

The installation of these new monitors was conducted for three years. In the first two years, the focus is on the upgrading of existing monitors in big cities, where existing manual stations in 113 cities were replaced with automatic stations. For some cities, additional monitors were installed since the city had expanded rapidly in the last few years. Cities in the Beijing-Tianjin-Hebei province, the Pearl River Delta, and the Yangtze River Delta, as well as the capital cities of each province, were assigned to upgrade/build monitor stations in 2012. All those monitor stations were connected with the environmental monitor network from the 1^{st} of January 2013 onward. In the second year, 30 environmental role model cities and 28 experimental cities from one province (Shandong) had set up local monitor stations and were connected to the network from the 1^{st} of January 2014. The largest expansion, which is also our research focus, took place the following year. From the 1^{st} of January 2015, all the remaining 177 cities were connected to the monitoring network. The location of these monitors is depicted in the map in Figure B2. After the final expansion, all prefecture-level cities had at least one air pollution monitor.

All monitor stations were built in the built-up area of each city.[7] The Ministry of Environmental Protection (MEP) provided instructions to provincial environmental departments for how the building/upgrading of monitors should be conducted. First, the minimum number of stations in each city is determined by the population of the city and the geographical size of the built-up area jointly. The detailed assignment criterion, which we use for identification in this paper, is presented in Table 4.1. Second, strict rules determine the placement of monitors to ensure that they are evenly distributed to cover the whole built-up area. These include applying a simulation method that takes surrounding buildings, traffic, and the direction of seasonal winds into account to make sure that the location of monitors is representative.

The provincial environmental departments were responsible for purchasing all the material, installing the monitor stations, testing the equipment, and overseeing daily operations. Once all equipment is put in place, the local environmental department is responsible for providing hourly measures of six pollutants to the Ministry of Environmental Protection, which in turn makes the information available to the public. The local governments, who face strong pressure from the central government to reduce pollution (see above), have the incentive to manipulate the data submitted to the environmental monitor network. The fact that local departments are in control of the monitors makes such manipulation feasible. A large number of media sources have reported that such manipulation has been extensive. Figure B9 shows an example of such manipulation.

Table 4.1: Monitor Assignment Criteria

Population (10,000)	Size of Buildup Area (sq. km)	Min # of Monitors
< 25	< 20	1
$25 - 50$	$20 - 50$	2
$50 - 100$	$50 - 100$	4
$100 - 200$	$100 - 200$	6
$200 - 300$	$200 - 400$	8
> 300	> 400	> 10

Sources: Technical regulation for selection of ambient air quality monitoring stations (the Ministry of Environmental Protection)

Retraction of Monitoring Stations Realizing that the data provided by local environmental protection departments might be manipulated, the Ministry of Environmental

[7] The built-up area is the urban area of the city as defined by the central government.

Protection (MEP) decided to contract the operation of monitor stations to private companies. According to official documents from the MEP, all the monitors were operated by private companies from the 1^{st} of November 2016. Monitors were procured through twelve contracts. Each contract was designed to involve monitors in different provinces spread out over the country, to make it difficult for rms to select into a given area. Six companies won the procurement, and each of them won two contracts. Importantly, after the operation of the monitors is taken over by the firms, all the operation costs are still paid by the Ministry of Environmental Protection (MEP) instead of the local government.

4.3. DATA

This section documents the main data used in this study. Section 4.3.1 describes the two sets of data that we use to measure air pollution: a satellite-based measure of the aerosol optical depth and data from the monitoring stations. After that, we describe the local air pollution enforcement records that we use. In Section 4.3.3, we discuss the additional data used, including satellite-based luminosity data, information about local leader characteristics, Baidu searches, and weather data. Finally, Section 4.3.4 discusses the summary statistics for our main sample.

4.3.1. AIR POLLUTION DATA

Satellite Data: Aerosol Optical Depth (AOD) Before the 2013-2015 expansion of the monitor system, most cities did not have any valid pollution monitoring. Even in cities covered by the old system, $PM_{2.5}$ information was not available. To address the problem of missing data before new monitors were built, we use Aerosol Optical Depth (AOD) provided by NASA to measure pollution. AOD measures the degree to which aerosols prevent the transmission of light by absorption or scattering of light, which is highly correlated with air quality (Gupta, Christopher, Wang, Gehrig, Lee, and Kumar, 2006). Monthly information on AOD is available at 0.1 by 0.1 degrees since 2000. In this project, we combine measures from the MODIS Aqua and Terra satellites to calculate the mean of AOD in a given month in a city. To deal with potential within city spillovers in pollution, we calculate this measure based on the whole prefecture-level polygon, as depicted in Figure B2. The mean of AOD in the data is 0.34, and the standard deviation is 0.23. To facilitate interpreting the magnitude, we use the natural logarithm of the AOD measure.

Air Pollution Data from Monitors Air pollution data is published online by the MEP. As discussed in the previous section, all 337 prefecture-level cities have at least one air pollution monitor since Jan 2015. Hourly and daily observations of SO_2, NO_2, CO,

PM_{10}, $PM_{2.5}$, and O_3 from 1,436 monitoring stations in 337 prefecture-level cities were recorded and published. Additionally, An Air Quality Index (AQI) was developed based on these six pollutants. The AQI scale ranges from 0 to 500. It is further divided into six ranges: $0-50$, $51-100$, $101-150$, $151-200$, $201-300$ and $301-500$. In public reports, these are categorized as good, moderate, unhealthy for sensitive groups, unhealthy, very unhealthy, and hazardous, respectively.

Aerosol Optical Depth (AOD) data has been used in various studies (Chen, Jin, Kumar, and Shi, 2013; Jia, 2017) to measure air pollution. Only a few studies have verified the correlation between AOD and the concentrations of fine particulate matter ($PM_{2.5}$). For example, Wang and Christopher (2003) finds that the correlation coecient between the monthly means of AOD and $PM_{2.5}$ is around 0.7 using data in Alabama in 2002. Using much more comprehensive data, Gupta, Christopher, Wang, Gehrig, Lee, and Kumar (2006) finds that the correlation varies between 0.14 and 0.6. Since we are studying a monitor expansion program, we take advantage of the ground measurements of $PM_{2.5}$ that are available after the expansion. We study the correlation between $PM_{2.5}/PM_{10}/AQI$ and the AOD data. The estimates are listed in Table 4.2, where we control for monitor fixed effects and time fixed effects in all regressions. Columns (1) show results for $PM_{2.5}$. The correlation is around 0.4 and that estimates, therefore, are comparable with Gupta, Christopher, Wang, Gehrig, Lee, and Kumar (2006). Columns (2) report the results for PM_{10} and show a similar pattern to that for $PM_{2.5}$. A similar pattern is also found for the correlation between the Air Quality Index (AQI) and the AOD, as shown in columns (3). The correlation between $PM_{2.5}$ and AOD is stronger then the one between PM_{10}/AQI and AOD. Not surprisingly, the correlation between AQI and AOD is the lowest since the Index also aggregates information from pollutants that are less correlated with AOD. All the evidence above supports that AOD is a suitable measure for local air pollution.

4.3.2. ENFORCEMENT RECORDS

We construct a new dataset on the enforcement activities of cities' environmental bureaus based on data provided by The Institute of Public & Environmental Affairs (IPO). IPO has, in turn, collected this data from all prefecture-level cities. To the best of our knowledge, this is the most reliable coverage of enforcement records in China. Our sample consists of the yearly number of records related to air pollution in each city between 2010 and 2017. Figure B4 in Appendix B provides an example illustrating the information contained in an enforcement record. Each record includes information about the violating firm, a description of the violation, a reference to the regulation that has been

Table 4.2: Validating Satellite Data

	(1)	(2)	(3)
	log(PM2.5)	log(PM10)	log(AQI)
log (AOD)	0.38***	0.32***	0.27***
	(0.016)	(0.016)	(0.011)
Monitor FE	Yes	Yes	Yes
Time FE	Yes	Yes	Yes
Observations	23831	23817	25232
R-squared	0.79	0.80	0.80

Notes: This table reports the AOD elasticity of $PM_{2.5}$. Each column is from a separate regression. All regressions control for fixed effects specific to monitor and time (month by year). Robust standard errors clustered on the monitor in parenthesis. *, **, *** indicates significance at the 10%, 5% and 1% level respectively.

violated, and the action taken by the local environmental bureau. These actions involve, e.g., suspending production, ordering upgrading of the equipment, and fines levied.

Based on the information provided about firm names, we link the enforcement records to the Environmental Survey and Reporting Database (ESR). The ESR database is a comprehensive environmental data set covering the majority of polluting industrial rms and their geographical location. These firms contribute to approximately 85% of China's total pollution. In this study, we keep only the ESR firms that are related to air pollution. Among a total of 22,000 records related to air pollution, 8,000 are enforcement actions against the firms in the ESR database.

4.3.3. ADDITIONAL DATA

Satellite Data: Luminosity Data To capture economic activity, we also rely on satellite-based measure of lights emitted during nighttime. The luminosity data has been used as a robust proxy for economic activity across and within countries since Henderson, Storeygard, and Weil (2012). We use data from the new Visible Infrared Imaging Radiometer Suite (VIIRS) instrument, which provides several improvements over the data of the Defense Meteorological Satellite Program (DMSP) (used in Henderson, Storeygard, and Weil (2012)). These include higher spatial resolution as well as a better ability to capture low-light emissions. Nighttime lights from VIIRS is available at the pixel-month level (roughly 0.86 square kilometers at the equator) since 2012. VIIRS light data is reported in Watts/m^2. Following Henderson, Storeygard, and Weil (2012), we calculate simple area averages of emitted light among the pixels within each county/district. This

is standard practice in the literature (Michalopoulos and Papaioannou, 2018).

Local Leader Characteristics Information on local officials is collected from the database compiled by Jiang (2018). The database contains extensive demographic and career information about over 4,000 key city, provincial and national leaders in China since the late 1990s. For each leader, the database provides standardized information about the time, place, organization, and rank of every job assignment listed in their curriculum vitae. Data is collected from government websites, yearbooks, and other trustworthy Internet sources. We merge this information of mayors to our panel data, with which we can calculate mayor's age in office for each month. This can, in turn, be used to infer the promotion incentives faced by the mayor, as discussed above.

Baidu Search Index To study the impact of new air pollution information, we collect data about local awareness of air pollution information from the Baidu Search Index. Similar to Google Trends (GT), Baidu Search Index provides a measurement of the search volume of a keyword in a given period from both computers and mobile devices. The Index is constructed by summing the weighted frequencies of all search queries for a specific keyword by city and by day. However, the exact algorithm of the Baidu Index is confidential and unknown to the public. Previous studies (Qin and Zhu, 2018; Barwick, Li, Rao, and Zahur, 2018) argue that the correlation between the Index and actual online search volume is linear. To match the frequency of our analysis on the air pollution data, we collect the monthly search volume from the Baidu Search Index of each city for the following keywords (in Chinese): air pollution, haze/smog, PM2.5, air mask, and air purifier.[8]

Weather Variables To control for local weather conditions, which are important determinants of the concentration of air pollution in prior work, we collect data from the China Meteorological Administration. This data uses the ANUSPLIN meteorological interpolation model to calculate average monthly temperature and precipitation at a 0.5-degree spatial resolution. We match this data to our prefecture-level cities to get a local measure of weather conditions.

4.3.4. SUMMARY STATISTICS

Summary statistics for the variables described above are presented in Table 4.3 for our main sample. This data covers the 177 cities that all installed monitors in 2015 and covers the period 2010-2017 for most variables.[9] The majority of the data is provided at the

[8] The Chinese translation of these five keywords are: 空气污染, 雾霾, PM2.5, 口罩, 空气净化器.
[9] VIIRS Light data is only available from 2012 onward.

monthly level.

Table 4.3: Summary Statistics

	Mean	Std. dev.	Observations
Panel A:			
Number of monitors	2.751	1.082	16,992
Urban Population (10,000)	31.56	22.16	16,992
Size of built-up area (km^2)	44.65	26.31	16,992
Aerosol Optical Depth	0.335	0.232	16,319
VIIRS Light	0.492	0.571	11,337
Panel B:			
Number of Enforcement	15.861	37.822	1,416
Number of Enforcement (ESR)	2.711	5.538	1,416
Panel C:			
Age of mayor	50.66	4.486	16,992
Precipitation (mm)	76.45	86.19	16,992
Mean Temperature	10.81	11.61	16,992

Notes: The table presents summary statistics for the samples used in our analyses. Data in Panel A and Panel C is provided at the monthly level. Enforcement data in Panel B is provided at the yearly level.

Our sample contains the 177 cities that installed new monitors in 2015. This focus is motivated by three main reasons. First, we do not want to mix cities that had monitors in the past (old monitors) with those that got a monitor for the first time (new monitors). The key reason for this is that the new information gained from an updated monitor is different since it captures recordings on a much wider set of pollutants. Second, cities with old monitors are dramatically different from cities with new monitors. In Table A2 we compare the descriptive statistics of cities with new monitors to those that had air pollution monitors before the reform. We see that the urban population and the size of the built-up area in cities with old monitors are 5-6 times larger. The size of the economy is also substantially different, as captured by the VIIRS light emission during night time. Finally, we exclude cities that received monitors at an earlier stage due to targeting related to proximity to industrial hubs (that faced different environmental policies) or environmental quality. This leaves us with our final sample in column (3).

4.4. MONITORING AND POLLUTION

This section presents our main analysis, i.e., estimating the impact of the number of monitors on a city's aggregate pollution. Section 4.4.1 describes the baseline empiri-

cal strategy (difference-in-differences) and reports the results. Section 4.4.2 discusses the robustness of the results of using an instrumental variable approach. We also consider different specifications and sample definitions in this section. Finally, Section 4.4.3 explores the third identification strategy (regression discontinuity) as an additional robustness check.

4.4.1. DIFFERENCE-IN-DIFFERENCES ESTIMATIONS

To study the effects of monitoring, we run regressions of the following form:

$$AOD_{it} = \delta_i + \gamma_t + \beta_1 Monitors_{it} + \delta X_{it} + \epsilon_{it}, \tag{4.1}$$

where the outcome AOD_{it} is the Aerosol Optical Depth for city i, in month t. $Monitors_{it}$ is the number of monitors in city i in month t and δ_i and γ_t are city and time fixed effects. X_{it} represent time varying characteristics of each city including: monthly precipitation, monthly average temperature and the age fixed effects of the mayor in office. The inclusion of weather controls is motivated by the fact that ambient pollution has been shown to be affected by local weather conditions in previous work (Schlenker and Walker, 2015; Barwick, Li, Rao, and Zahur, 2018). To ensure that the number of monitors installed does not depend on the motivation of the local leaders, we check the robustness of our results to controlling for mayor characteristics. The identification of a causal effect relies on the common trend assumption.

Table 4.4 reports the main results from the estimation of Equation 4.1, showing the effect of monitoring on air pollution measured by the logarithm of monthly Aerosol Optical Depth (*AOD*). We look at all cities that built monitor stations in 2015. All four columns use the difference-in-differences strategy by comparing the change in pollution before and after the policy between cities that installed a different number of monitors. We only control for city fixed effects and time fixed effects in the first column. We add controls for time-varying weather conditions in column 2. And in the third column, we also control for the characteristics of the mayor in office. In the last column, we add controls for region-specific linear trends.[10] These are included to ensure that results are not driven by a different number of monitors installed in cities that face different targets of pollution reduction. The standard errors in the regressions are clustered at the city level. The estimates consistently show that one more monitor leads to a 2-3 percent decrease in air pollution. A causal interpretation requires the common trend assumption that cities that installed a different number of monitors would have had similar pollution

[10] The regions consist of several provinces and are defined according to the Table A1.

trends in the absence of monitors.

Table 4.4: Impact of Monitoring on Pollution

	(1)	(2)	(3)	(4)
Outcome:	AOD			
Monitors	-0.027***	-0.027***	-0.025***	-0.026***
	(0.0037)	(0.0037)	(0.0037)	(0.0038)
City FE	Yes	Yes	Yes	Yes
Time FE	Yes	Yes	Yes	Yes
Weather Controls	No	Yes	Yes	Yes
Mayor Controls	No	No	Yes	Yes
Region Trend	No	No	No	Yes
Observations	16319	16319	16319	16319

Notes: This table reports estimates of one additional monitor effects on air pollution. Each column is from a separate regression specified in Equation 4.1. All regressions control for fixed effects specific to city and time (month by year). Weather controls include precipitation and average temperature at month level. Mayor controls include age fixed effects of the mayor in office. Robust standard errors clustered on the city in parenthesis. *, **, *** indicates significance at the 10%, 5% and 1% level respectively.

To test for differential pre-trends, we conduct event studies based on the specifications used above - setting the average pollution in the year before monitors were installed as the baseline. Figure 4.1a presents graphical evidence for our event study, where we look at the same dependent variable, *AOD*, as in our baseline model. We also control for time-varying weather conditions and the characteristics of the mayor in office. In Figure 4.1a, there is no evidence of differential trends leading up to the intervention. In the year of adoption, we see a substantial drop in pollution for cities with more monitors compared to cities with fewer.[11] These effects are even stronger in the second and third years.

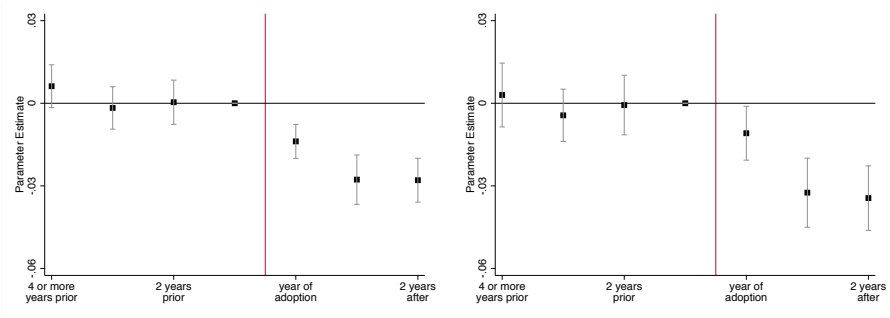
4.4.2. ROBUSTNESS: SPECIFICATIONS AND SAMPLE DENITIONS

In this section, we explore some additional specification checks to make sure that our estimates from the previous section can be interpreted as the causal effects of the monitor program. Table 4.5 presents these additional results.

Our baseline model using OLS might lead to biased results if the number of monitors installed is endogenous. The estimates would be biased if, for example, cities that expected lower pollution in the future installed a larger number of monitors. To address

[11] Note that monitors are operational from the first of January, so all periods in the year of adoption are treated.

Figure 4.1: Event Study



(a) DiD (b) DiD + IV

Notes: The figures present the results from an event study estimating the effects of monitoring using two different specifications (DiD, DiD+IV). The figures plot the coefficients from the event study. Capped spikes represent 95 percent confidence intervals. Both estimates control city fixed effects, time fixed effect, weather conditions and the characteristics of the mayor. See the text for further details.

these concerns, we propose to use the minimal number of monitors set by the Ministry of Environmental Protection (MEP) as an exogenous source of variation. More specifically, we explore the variation created by the rule which determines the minimal number of monitors in each city. As outlined in Table 4.1, larger cities (measured by population and the geographical size of the built-up area) were requested to build more monitor stations. Therefore, the first specification uses the minimum number of monitors according to the rule as the instrument for the actual number of monitors ($Monitors_{it}$). Therefore, the causal effect relies on the common trend assumption between cities with different sizes of the population and the built-up area.

Panel A in Table 4.5 displays the instrumental variable estimates. As in our baseline DiD model, we gradually add time-varying controls: weather conditions, the mayor's age in office, and regional specific linear trends. Results are very similar to what we find in the baseline model (point estimates in the most conservative specification are the same), are robust across specification, but are somewhat less precisely estimated than in the baseline DiD. To test for the common trends assumption, we implement an event study specification for this strategy. Figure 4.1b presents the results from this exercise and again shows no evidence of pre-trends and point estimates and timing of the effect similar to the results from the baseline specification.

In Panel B of Table 4.5, we drop data from both Xinjiang and Tibet since the area covered by each city in these two provinces is much larger than for the rest of the country.

Table 4.5: Robustness

	(1)	(2)	(3)	(4)
Outcome:			AOD	
Panel A: Instrumental Variable				
Monitors	-0.030***	-0.030***	-0.028***	-0.026***
	(0.0049)	(0.0049)	(0.0050)	(0.0053)
F-Stat of 1^{st} stage	161.5	160.7	151.1	147.8
Observations	16319	16319	16319	16319
Panel B: Sample Restriction				
Monitors	-0.023***	-0.022***	-0.022***	-0.024***
	(0.0038)	(0.0038)	(0.0037)	(0.0037)
Observations	14625	14625	14625	14625
City FE	Yes	Yes	Yes	Yes
Time FE	Yes	Yes	Yes	Yes
Weather Controls	No	Yes	Yes	Yes
Mayor Controls	No	No	Yes	Yes
Region Trend	No	No	No	Yes

Notes: This table reports estimates of one additional monitor effects on air pollution. Each column in each panel is from a separate regression specified in Equation 4.1. Panel A uses the minimal number of monitors set by MEP as the instrumental variable. Panel B drops sample from both Xinjiang and Tibet. All regressions control for fixed effects specific to city and time (month by year). Weather controls include precipitation and average temperature at month level. Mayor controls include age fixed effects of the mayor in office. Robust standard errors clustered on the city in parenthesis. *, **, *** indicates significance at the 10%, 5% and 1% level respectively.

The OLS estimates using the restricted sample is slightly smaller if we do not include any time-varying controls. But when we consider all time-varying controls, the results are very similar to what we find in the DiD model. All the evidence above shows that our results are robust to different specifications and sample selections.

4.4.3. ROBUSTNESS: EVIDENCE FROM RD

While all results from the above implemented empirical strategies and specification checks suggest that we capture the causal effect of the number of monitors on pollution, a potential remaining concern is that cities of different sizes face different incentives to reduce pollution after the reform. To address this concern, we explore the variation caused by discontinuities in the number of monitor stations at city size and population cutoffs (as documented in Table 4.1).

The strategy is essentially a fuzzy regression discontinuity design, where the identifi-

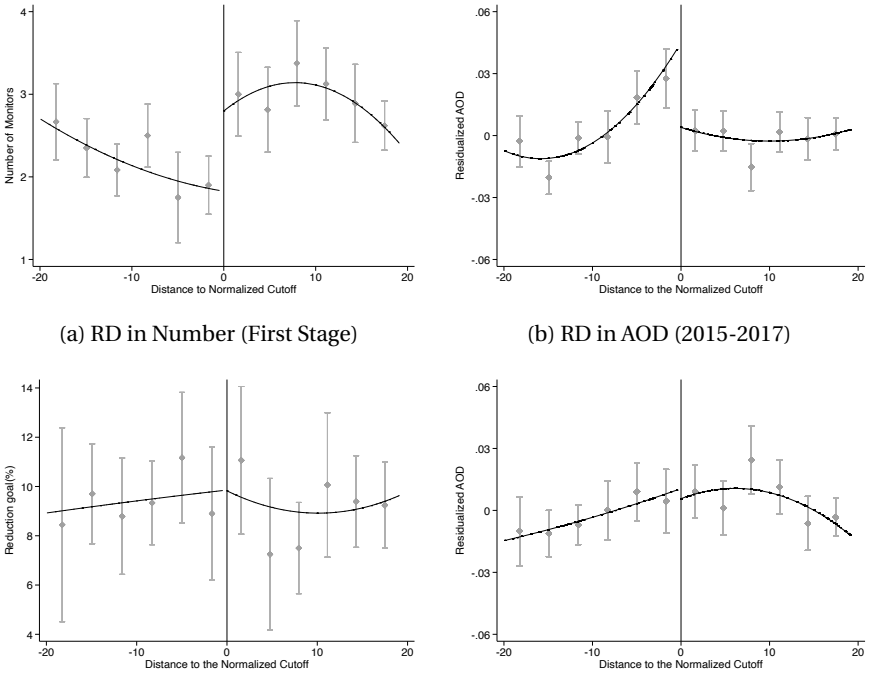
cation relies on the assumption that all other city characteristics change smoothly at the cutoffs. Table 4.1 shows the criteria determining the minimum number of monitors for each city. Compared to the standard regression discontinuity design using one running variable and cutoff, we have two running variables and multiple cutoffs. In practice, we notice that the population's size is not a good predictor of the realized number of monitors. Hence, we only use the size of the buildup area as the running variable. And to improve the statistical power of regression discontinuity analysis, especially that only a limited number of cities are close to the cutoffs, we pool all observations together, regardless of the number of cities around the cutoff, and make inferences as in a standard RD design with a single cutoff. As documented in Table 4.1, there are 5 cutoffs in total. However, our sample only contains middle and small-sized cities, and there are only a sufficient number of cities on both sides of the threshold for the first two cutoffs. Among the 177 cities, only three cities have a population larger than 1 million or a geographical size of the built-up area that is larger than 100 km^2. We, therefore, focus on the first two cutoffs.

First-stage estimates for these two cutoffs are reported in Figure 4.2a, where we draw scatters and quadratic fits of the number of monitors in each city. Cutoff fixed effects are absorbed before plotting the regression discontinuities. These graphs allow us to see the first-stage regression discontinuity estimates. The number of monitors exhibits a sharp jump when moving from the left to the right of the threshold. The first-stage estimates show that cities just above the threshold have installed approximately 1.5 additional monitors. Figure 4.2b shows the reduced-form estimates on AOD, where we draw scatters and quadratic fits of the AOD values. We see clear jumps in AOD when moving from the left to the right of the threshold.

Table 4.6 quantifies the graphical findings in Figure 4.2. First three columns report the RD regression using different kernel weighting methods. The discontinuities are estimated using local linear regressions. RD regressions control for cutoff fixed effects and baseline average AOD. The last column reports the estimate from a Difference-in-Discontinuities regression proposed by Grembi, Nannicini, and Troiano (2016), which also exploit the longitudinal nature of the data. The idea is essentially to estimate the baseline non-parametric RD model while interacting every term with a dummy variables indicating "post-2015".[12] In all different specifications, we cluster standard errors on the city. The evidence is very clear that cities have substantially lower satellites measured AOD than those slightly smaller cities whose size of the buildup area are just above the cutoff. The drop in AOD across the cutoff is related to an increase in enforcement.

[12] More details about the Difference-in-Discontinuities strategy can be found in the Appendix C.

Figure 4.2: Regression Discontinuity Plots



(a) RD in Number (First Stage)

(b) RD in AOD (2015-2017)

(c) RD in Reduction Targets

(d) RD in AOD (2012-2014)

Notes: Cutoff fixed effect are absorbed before plotting the regression discontinuities. Baseline AOD are also controlled for Panel 4.2b and 4.2d. Robust standard errors clustered at the city level. Error spikes represent 95 percent confidence intervals.

Although less precise, all regression discontinuity estimates are very close to the counterparts using the difference-in-differences specification. The results are also robust to alternative bandwidths, as shown in the Figure 4.3.

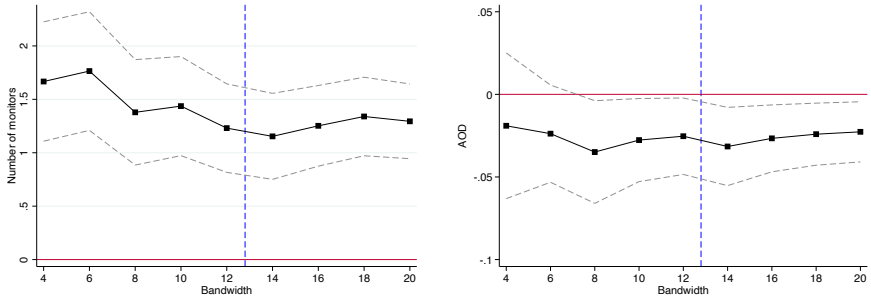
The above estimation results rest on the standard assumption that there is no manipulation of the running variable and that other characteristics of cities are smooth at the thresholds. If mayors were able to manipulate population size and the size of the built-up area and sort below the threshold to avoid an additional monitor, our estimates would still suffer from the selection bias. Figure B5 is reassuring about the absence of manipulation, as there is no jump in the distribution at each threshold. To test the second assumption, we study the main threat to this identification strategy - that cities with a different number of monitors face different pollution reduction targets. We look at targets for cities close to the thresholds using the same cross-sectional specification as used above to estimate the first stage impact on the number of monitors. Figures 4.2c re-

Table 4.6: Estimates from RD

	(1)	(2)	(3)	(4)
		RD		Diff-in-Disc
Monitors	-0.024**	-0.027*	-0.028**	-0.025*
	(0.012)	(0.014)	(0.014)	(0.014)
Obs.	3960	2376	2448	9216
Bandwidth	12.8	8.57	8.91	12.8
First stage	1.50***	1.61***	1.79***	1.50***
	(0.26)	(0.22)	(0.25)	(0.26)
Kernel	Uniform	Epanechnikov	Triangle	Uniform

Notes: This table reports estimates of one additional monitor's effects using a regression design. First three columns report the RD regression using different kernel weighting methods. The discontinuities at the normalized cutoff are estimated using local linear regressions and MSE-optimal bandwidth proposed by Calonico, Cattaneo, and Titiunik (2014) for repective kernel weighting methods. RD regressions control for cutoff fixed effects and baseline average AOD. The last column reports the diff-in-dist regression proposed by Grembi, Nannicini, and Troiano (2016). Robust standard errors clustered on the city in parenthesis. *, **, *** indicates significance at the 10%, 5% and 1% level respectively.

Figure 4.3: Alternative Bandwidths



(a) RD in Number (First-stage) (b) RD in AOD

Notes: Vertical axis: RD coefficients. Horizontal axis: bandwidth used to estimate the reported RD coefficients. The blue dashed line marks the optimal bandwidth (12.8) of AOD selected following Calonico, Cattaneo, and Titiunik (2014).

port the results from this exercise and show that pollution reductions targets are smooth around the thresholds. This suggests that differential pollution reduction targets do not drive our results. As additional checks, we present RD plots (Figures 4.2d) of AOD for pre-policy periods (2012-2014). To the contrary of the post-policy periods (2015-2017), we see no jumps at the threshold of the normalized the running variable. If there is any jump in other characteristics of cities at the thresholds, the violation of the second assumption would be reflect in these two figures.

The regression discontinuity estimates corroborate the results from the main panel specification, indicating large pollution drop following the set up of the air quality monitor. The strength of the regression discontinuity approach is its reliance on a few assumptions for causal inference. But the power of the test is limited, as well as the restriction of the sample close to the threshold. We, therefore, focus on the panel specifications in the section on treatment heterogeneity below.

4.5. ENFORCEMENT OF ENVIRONMENTAL REGULATIONS

In previous sections, we have documented the causal relationship between the decline in air pollution and the number of air pollution monitors. We hypothesize that this reduction is driven by local governments being stricter in the enforcement of air pollution-related regulations in the areas close to a monitor. An additional monitor increases coverage of total city pollution and therefore causes a decrease in air pollution.

To explicitly test this hypothesis, we collect data on the yearly number of the enforcement records of each city in our sample. For a total of 177 cities, 22,000 official enforcement records were recorded for the period from 2010 to 2017. A large share (75%) of the enforcement measures were taken after air pollution monitors were introduced. The identification strategy is the same as our main empirical strategy in section 4.4.1. The only difference is that we control for year fixed effects instead of month fixed effects since the data is now provided at this more aggregate level. Table 4.7 shows estimates using the difference-in-differences strategy. The first column reports the effects of one additional monitor on total enforcement. We see that one additional monitor leads to a 20% increase in the number of enforcement activities related to air pollution. The event study graph B7 in Appendix B verifies that the common trend assumption is indeed valid in this case as well.

To ensure that air pollution monitors drive the effects on air pollution and enforcement practices, we zoom in on the data by looking at the geographical pattern of enforcement practices. As most of the monitors were placed in the built-up area of each city, emission from plants located far away from urban areas will have less impact on readings of these monitors. Therefore, we expect that firms would be treated differently based on their location relative to the monitors. Firms that are located close to air pollution monitors might be subjected to tighter environmental enforcement compared to firms located far away. The literature has documented this phenomenon in some other settings, such as tax collection (Liu, 2017). To test this hypothesis in our setting, we link the enforcement data set with the Environmental Survey and Reporting Database (ESR), which provides us with location information for these firms. We aggregate the total num-

Table 4.7: Impact of Monitoring on Enforcement Practices

	(1)	(2)	(3)	(4)	(5)	(6)
	Total	ESR Firm	0-20 km	20-40 km	40-60 km	60- km
Outcome:	log(# Enforcement)					
Monitors	0.21***	0.16***	0.16***	0.034	0.038	0.026
	(0.051)	(0.042)	(0.037)	(0.025)	(0.024)	(0.030)
City FE	Yes	Yes	Yes	Yes	Yes	Yes
Time FE	Yes	Yes	Yes	Yes	Yes	Yes
Observations	1416	1416	1416	1416	1416	1416

Notes: This table reports estimates of one additional monitor effects on city-level environmental enforcement. Each column is from a separate regression. All regressions control for fixed effects specific to city and year. Column (1) reports the estimate using all firms. Column (2) reports the estimate using firms in the Environmental Survey and Reporting Database (ESR). Column (3) to (6) report estimates using ESR firms in four different distance bands. Robust standard errors clustered on the city in parenthesis. *, **, *** indicates significance at the 10%, 5% and 1% level respectively.

ber of enforcement activities against ESR firms in each city. In column 2 of Table 4.7, we look at the effects of an additional monitor on enforcement that targets ESR firms. The DiD estimate shows that one additional monitor increases enforcement by 16% , which is slightly lower than the increase in total enforcement.

We geo-coded all firms in the ESR database and calculated the distance of each firm to the city center. Records of enforcement practices for all firms in the ESR database are aggregated using four distance bands: 0-20 kilometers from the city center, 20-40 kilometers, 40-60 kilometers, and over 60 kilometers. The same model is then estimated separately for each group, and the estimates are displayed in columns 3-6 of Table 4.7. We notice that one additional monitor increases enforcement practices by 16% for firms located within 20km from the city center, while enforcement records barely increase for firms located further away from the city center. The steep gradient over distance in enforcement supports our hypothesis that local governments do put stricter standards on environmental regulations in areas close to a monitor.

One caveat in the above analyses is that the density of firms is higher in areas close to the city center.[13] Therefore, it might be the case that the larger increase in areas closer to built-up areas is simply due to the larger number of firms in this area. To address this concern, we construct a balanced panel of firms, where we examine the effects of an additional monitor within 20km. The estimates are shown in Table 4.8. In the first column,

[13] The distribution of distance between firms in the ESR database with the city center is displayed in Figure B6.

we control for the county (lower admin-level than city) fixed effects and year fixed effects. The estimates show that a firm is subject to around 2.4 percent more enforcement if one additional monitor is located within 20km. In the second column, we control for firm fixed effects and document similar estimates. In the third column, we also control whether the firm is located within 20km from the city's geometric center. We find that the estimate remains similar, and there is no effect of locating close to the city center. We rule out an alternative explanation that the increase of enforcement is targeted at firms in the city center. The fourth column displays the results of using a non-parametric approach concerning distance to estimate treatment effects. The effects of monitors more than 20km away are all much smaller and not significantly different from 0. The estimates again verify that the enforcement practices are only centered around air pollution monitors. In the last column, we use the continuous measure, $log(distance)$ to the closest monitor, as the measure of distance. Again, we see that firms located closer to the monitor have more enforcement records after 2015. As the identification strategy is again difference-in-differences, we also show the event study graph B7 in Appendix B to verify the common trend assumption. All the above evidence shows that local officials have increased the enforcement of environmental regulations in areas close to air pollution monitoring stations. These results are also consistent with the evidence documented in He, Wang, and Zhang (2020).

4.6. MECHANISMS

4.6.1. PROMOTION INCENTIVES

Our empirical analyses have shown that air pollution and the enforcement of regulations are affected by air quality monitoring. We rationalize our findings that monitor readings are politically important, and local officials, therefore, have incentives to engage in stricter enforcement against firms located close to the monitor. In this section, we investigate this proposed mechanism more directly by a series of heterogeneity analyses by promotion incentives of local officials. The first variation we exploit is the age of the mayor in each city. Mayors of prefecture-level cities are required to retire at age 60 at the same time as these officials are supposed to serve for at least three years in a post. Therefore, city officials above the age of 57 face a discontinuously lower probability of being promoted and, therefore, weaker performance incentives.

We use this age cutoff to capture officials' promotion incentives. More specifically, we collect data on all mayors in office the year before the introduction of monitors and cal-

Table 4.8: Impact of monitoring on Enforcement Practices

	(1)	(2)	(3)	(4)	(5)
Outcome:	log(# Enforcement)				
# monitors within 20km	0.024***	0.023***	0.032*	0.021*	
	(0.0088)	(0.0082)	(0.016)	(0.011)	
Within 20km from City Center			-0.028		
			(0.051)		
# monitors in 20-40km				-0.0031	
				(0.012)	
# monitors in 40-60km				-0.010	
				(0.013)	
log(distance) to					-0.021*
the closest monitor					(0.012)
County FE	Yes	No	Yes	Yes	Yes
Firm FE	No	Yes	No	No	No
Year FE	Yes	Yes	Yes	Yes	Yes
Observations	8808	8808	8808	8808	8808

Notes: This table reports estimates of one additional monitor effects on firm-level environmental enforcement. Each column in each panel is from a separate regression. All regressions control for fixed effects specific year. Column (1), (3), (4) and (5) control for fixed effects specific to county. Column (2) controls for fixed effects specific to firm. Robust standard errors clustered on the city in parenthesis. *, **, *** indicates significance at the 10%, 5% and 1% level respectively.

culate their age at the 13^{th} National People's Congress, which was held in March 2018.[14] Hence, what will matter for promotion is the age of an official at the time of the congress. If local officials' incentive is complementary to the new monitoring system in reducing pollution, we would expect smaller effects of monitoring when cities have mayors above 57 years of age at the time of the congress. Mayors who are not facing promotion incentives are arguably less likely to make efforts to reduce pollution, such as enforcing stricter pollution regulations (e.g., car pollution standards, stubble burning), monitoring environmental compliance of local rms and choosing whether to subsidize cleaner energy.

To test our hypothesis about promotion incentives formally, we employ specification 4.1 and further add an interaction term between $Monitor_{it}$ and a dummy variable for whether the mayor is older than 57. The regression results are reported in Table 4.9,

[14] The National Congress, which is held every five years, determines political cycles in China. As documented in Xi, Yao, and Zhang (2018), the average probability of promotion for a city official in the last year of a political cycle is nearly three times that of the first year in a cycle.

where Panel A displays the results on air pollution, and Panel B displays the results on enforcement. In the first column, we use the full sample as in our main analyses in Section 4.4.1. The coefficients for $Monitors_{it}$ are also almost identical to the ones obtained in our previous analyses. The interaction term's coefficient has the opposite sign, which means that the effect is smaller in cities with a retiring mayor. The magnitude in Panel A is also close to the effect of having one additional monitor – suggesting that there is no effect of having one more monitor when there are no promotion incentives. The magnitude of the interaction term in Panel B is about half of the effect on the enforcement of having one more monitor – suggesting that retiring mayors enforce significantly less. The full sample includes mayors of all different ages, which contains much younger mayors. Since mayors' work experience might confound our analysis, we use a similar idea as in the RD design and restrict the sample by mayors' age to check the robustness of our analysis. We gradually reduce the age bandwidth in columns (2)-(4), and the coefficients are only slightly smaller and remain significant when we decrease the bandwidth. In graphs B8a and B8b, we plot the differential effects (i.e. the interaction terms) of an additional monitor on both pollution and enforcement by the age of the mayor at the time of the congress. We normalize the effect to 0 for cities with a mayor who would be 50 years old in 2018. A distinctive feature of both graphs is that the effects are not distinguishable from 0 if the mayor is younger than 58. At age 58, we see a substantial jump of the estimates in both graphs. Therefore, it is most likely that the results we show are driven by the promotion incentives of mayors instead of by other characteristics such as mayors' work experience. We conclude from the analyses that pre-existing promotion incentives are key for monitoring to have an impact on air pollution.

4.6.2. CITIZEN ENGAGEMENT

A growing literature (Chen, Pan, and Xu, 2016; Meng, Pan, and Yang, 2017) has found that authoritarian regimes are responsive to societal actors. For example, Chen, Pan, and Xu (2016) find that approximately one-third of local county governments in China respond to citizen demands expressed online. Local governments are more responsive to the threats of collective action, which are often seen in environment-related issues.[15] Coupled with the fact that the awareness of the health impact of $PM_{2.5}$ has increased since 2010 (Barwick, Li, Rao, and Zahur, 2018), we believe that another possible mechanism is that information from additional monitors increases citizen awareness about pollution levels and their engagement. Such engagement could potentially further drive

[15] See `https://www.bbc.com/zhongwen/simp/china/2015/06/150624_shanghai_chemicalplant` as an example.

Table 4.9: Promotion Incentives

	(1)	(2)	(3)	(4)
Age bandwidth:	Full	±10 Years	±7 Years	±5 Years
Panel A:		AOD		
Monitors	-0.026***	-0.026***	-0.027***	-0.022***
	(0.0035)	(0.0036)	(0.0038)	(0.0044)
Monitors × Above 57	0.019***	0.018***	0.017***	0.021***
	(0.0046)	(0.0048)	(0.0048)	(0.0030)
City FE	Yes	Yes	Yes	Yes
Time FE	Yes	Yes	Yes	Yes
Observations	16319	15208	14000	11084
Panel B:		log(# Enforcement)		
Monitors	0.21***	0.17***	0.18***	0.16**
	(0.051)	(0.053)	(0.057)	(0.064)
Monitors × Above 57	-0.11**	-0.10**	-0.098*	-0.10*
	(0.051)	(0.051)	(0.051)	(0.055)
City FE	Yes	Yes	Yes	Yes
Time FE	Yes	Yes	Yes	Yes
Observations	1416	1280	1176	936

Notes: This table reports heterogeneous effects of one additional monitor on air pollution (Panel A) and environmental enforcement (Panel B). Each column in each panel is from a separate regression. All regressions control for fixed effects specific to city and time (month by year in Panel A and year in Panel B). Robust standard errors clustered on the city in parenthesis. *, **, *** indicates significance at the 10%, 5% and 1% level respectively.

up efforts to protect the environment after the air pollution information is released. To formally test this hypothesis, we estimate Equation 4.1 to identify the causal effects of monitoring on online search behavior. Table 4.10 shows the estimates for five pollution-related keywords. Column 1 documents that search on air pollution increases by around 1% for one additional monitor station. However, all other estimates are not significantly different from 0. We believe these effects are likely too small to explain the decrease in air pollution that we document in the previous section – especially considering that real action or engagement would probably be lower than the increase in online searches.

Table 4.10: Impact of monitoring on Online Searches

	(1)	(2)	(3)	(4)	(5)
Key words	air pollution	haze/smog	PM2.5	air mask	air purifier
Outcome:	log(Search Index)				
Monitors	0.0095***	0.026	0.00010	0.0074	0.023
	(0.0033)	(0.017)	(0.00060)	(0.0098)	(0.017)
City FE	Yes	Yes	Yes	Yes	Yes
Time FE	Yes	Yes	Yes	Yes	Yes
Observations	14610	14610	14610	14610	14610

Notes: This table reports estimates of one additional monitor effects on online search behavior. Each column is from a separate regression looking at different key words. All regressions control for fixed effects specific to city and year. Robust standard errors clustered on the city in parenthesis. *, **, *** indicates significance at the 10%, 5% and 1% level respectively.

4.6.3. QUALITY OF INFORMATION

Although providing incentives for performance is a common approach to deal with the principal-agent problem, it has long been recognized that high powered incentives can also distort the type of effort exerted or even encourage various harmful activities focused on improving indicators of performance. Manipulating indicators of performance is a typical strategy that has been documented in a series of studies. In this section, we study whether the party in control of the monitors matter for the quality of the information provided and whether such quality improvements can, in turn, strengthen accountability and government performance.

Several media sources reported on extensive manipulation of the pollution data by local government officials. Such manipulation took many different forms – ranging from directly adjusting the numbers to spraying the monitors with water, as shown in Figure B9 in the Appendix. Following this reporting, the central government decided to reassign the control of monitors to external parties, as documented above. In this section, we take advantage of this retraction policy to see whether increasing the cost of manipulation is an effective way to improve monitoring, reduce manipulation, and through that, enforce environmental policy.

In order to identify the impact of the reassignment on the quality of information, we study how the AOD elasticity of $PM_{2.5}$ changes when the way information is provided

changes (I_t). More specifically, we estimate:

$$log(PM_{2.5})_{it} = \delta_i + \gamma_t + \beta_1 log(AOD)_{it} + \beta_2 I_t + \beta_3 log(AOD)_{it} \times I_t + \epsilon_{it}, \quad (4.2)$$

where $log(PM_{2.5})_{it}$ is the natural logarithm of monthly average concentrations of $PM_{2.5}$ reported from monitor i at time t. δ_i and γ_t represent fixed effects for monitors and time. The variable $log(AOD)_{it}$ captures the Aerosol Optical Depth for the pixel covering monitor i at time t.[16] I_t is a dummy variable indicating whether the data is reported after the retraction. Therefore, the main coefficient of interest is β_3. If the retraction increased the quality of information, we would expect that AOD and $PM_{2.5}$ measures are more aligned and therefore that $\beta_3 > 0$.

The results from estimating Equation 4.2 are reported in the first column of Table 4.11. We find a positive estimate for the interaction term. This shows that the elasticity is 0.1 larger (corresponding to a 29% increase in the elasticity) after the third party takes over the monitoring stations – suggesting that there is less manipulation and that monitors provide higher quality information. To further understand the mechanism behind this, we exploit heterogeneity in the relationship between these private firms and the local governments. We identify such relationships by exploiting whether the firm in control of a monitor is located in the same province. A local firm might be subject to political influence by the local government or have political connections to local officials, allowing manipulation to continue. The results are shown in columns (2) and (3) of Table A3 in Appendix A. In line with such an interpretation, we find that non-local firms drive the entire increase in the elasticity (i.e., the decrease in manipulation). The coefficient for local firms is negative, which suggests that they might collude with the government and help local officials to manipulate the data. The takeaway of the above findings is that a third-party may not be sufficient for preventing manipulation if they are not sufficiently insulated from government influence.

The next excise we carry out is to check whether local governments exert more effort to decrease pollution after the cost of manipulation has increased. The results are reported in the second and the third column of Table 4.11. The second column shows a 1.1 percent stronger decrease in pollution measured using satellite data after the retraction. Compared to the baseline estimates, one more monitor increased by one third after the operation was handed over to private firms. The larger drop in air pollution due to an additional monitor is related to a larger increase in enforcement efforts. As shown in column 3, the effect of one more monitor is 44% larger than in the baseline. These two

[16] To deal with the fact that data is sometimes missing for the pixel just above the monitor, e.g. because of cloud coverage, the value for missing pixels is interpolated from surrounding cells.

Table 4.11: Retraction and Manipulation

	(1)	(2)	(3)
Outcome:	log(PM2.5)	AOD	log(# Enforcement)
log(AOD)	0.35***		
	(0.023)		
log(AOD) × Post Retraction	0.10***		
	(0.029)		
Monitors		-0.022***	0.18***
		0.0038	0.047
Monitors × Post Retraction		-0.011***	0.082*
		0.0024	0.049
Monitor FE	Yes	No	No
City FE	No	Yes	Yes
Time FE	Yes	Yes	Yes
Observations	17957	15938	1384

Notes: Each column is from a separate regression. Column (1) reports the change of AOD elasticity of $PM_{2.5}$ after the third party takes over the monitoring stations. Column (2) and (3) reports heterogeneous effects of one additional monitor on air pollution and environmental enforcement. Robust standard errors clustered on the city in parenthesis. *, **, *** indicates significance at the 10%, 5% and 1% level respectively.

pieces of evidence support our hypothesis that if it is difficult for local governments to manipulate data, they exert more effort into enforcing environmental regulations.[17]

4.7. CONCLUDING REMARKS

This study uses the introduction of a nationwide pollution monitoring program in China to investigate the impact of installing additional monitors on both pollution concentration and government enforcement of pollution regulations. We show that installing an additional pollution monitor reduces actual pollution by 2-3% as measured by satellite data. This reduction in air pollution is arguably driven by a 20% increase in the number of regulatory enforcement activities. Exploiting geo-referenced enforcement records, we find that the increase in enforcement is driven by firms located within 20km from a monitor, whereas firms beyond 20km face no change in enforcement.

An examination of possible mechanisms shows that the career concerns of the local officials can explain a big part of our findings. This suggests that accurate monitoring of environmental quality can strengthen accountability and improve policy outcomes.

[17] He, Wang, and Zhang (2020) also documented a similar phenomenon around water monitor stations.

On the contrary, we find little evidence that citizen engagement is affected by additional monitors. Finally, we document that when monitors are insulated from government influence and external parties are responsible for the provision of information, the effect of an additional monitor on enforcement and reductions of pollution are both stronger.

Our findings are broadly related to a literature looking into the principal-agent problem in bureaucracies. Local officials have the incentive to game the system by shirking or manipulating the data, especially when local information is costly to verify. Our paper provides new empirical results on local officials' efforts and the environmental consequences of introducing high-quality information. This finding relates to a large theoretical literature that tries to design optimal incentive and information structures.

Due to the recency of the reform that we study, the availability of economic data is limited. Therefore, we are not able to say much about the impact of monitoring on economic outcomes (such as local economic growth or the exit/entry of firms). Given a strong emphasis on economic performance by many developing countries, understanding whether improved monitoring and enforcement have implications for economic performance and whether there is a trade-off between improved environmental and economic outcomes is a promising question for future research.

A. ADDITIONAL TABLES

Table A1: Targets by Province

Targeted Pollutants	Target	Provinces
$PM_{2.5}$	-25%	Beijing, Tianjin and Hebei
$PM_{2.5}$	-20%	Shagxi, Shandong, Shanghai, Jiangsu, Zhejiang
$PM_{2.5}$	-15%	Guangdong, Chongqing
$PM_{2.5}$	-10%	Inner mongolia
PM_{10}	-15%	Henan, Shannxi, Qinghai, Xinjiang
PM_{10}	-12%	Gnasu, Hubei
PM_{10}	-10%	Sichuan, Liangning, Jilin, Hunan, Anhui, Ningxia
PM_{10}	-5%	Guangxi, Fujian, Jiangxi, Guizhou, Heilongjiang
PM_{10}	0%	Hainan, Tibet, Yunnan

Sources: The Ministry of Environmental Protection (MEP)

Table A2: Summary Statistics

	(1)	(2)	(3)
	Old Monitors	New Monitors	2015 Sample
Outcomes			
AOD (Aqua)	0.43	0.35	0.32
	(0.24)	(0.24)	(0.23)
AOD (Terra)	0.46	0.38	0.35
	(0.25)	(0.25)	(0.24)
AOD	0.45	0.37	0.34
	(0.24)	(0.24)	(0.23)
VIIRS light	2.28	0.69	0.49
	(2.85)	(1.10)	(0.57)
Controls			
Number of Monitors	6.26	3.24	3.12
	(2.75)	(1.20)	(1.26)
Urban Population	195.1	39.5	33.8
	(303.7)	(26.9)	(21.0)
Size of Buildup Area	257.7	60.2	46.8
	(345.1)	(74.1)	(27.2)
New City	0	1	1
Number of observations	10848	21600	16992

Notes: Author's tabulations.

Table A3: Retraction and Manipulation

	(1)	(2)	(3)
Outcome:		log(PM2.5)	
log(AOD)	0.35***	0.36***	0.36***
	(0.023)	(0.023)	(0.023)
log(AOD) × Post Retraction	0.10***		
	(0.029)		
log(AOD) × Post Retraction × Non-Local Firm		0.11***	
		(0.029)	
log(AOD) × Post Retraction × Local Firm		-0.14*	-0.14*
		(0.076)	(0.076)
log(AOD) × Post Retraction × Never Local			0.14***
			(0.033)
log(AOD) × Post Retraction × Local Elsewhere			0.036
			(0.041)
Monitor FE	Yes	Yes	Yes
Time FE	Yes	Yes	Yes
Observations	17957	17957	17957

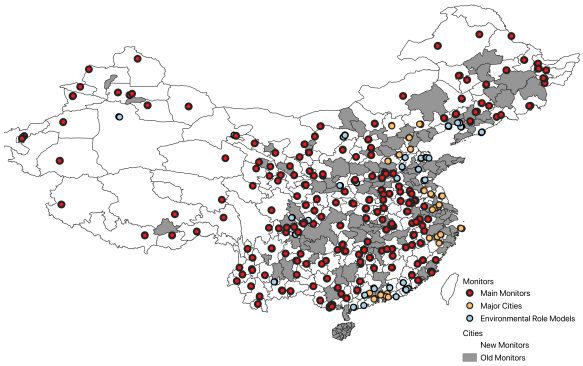
Notes: Each column is from a separate regression. All regressions control for fixed effects specific to monitor and time (month by year). Robust standard errors clustered on the city in parenthesis. , , indicates significance at the 10%, 5% and 1% level respectively.

B. ADDITIONAL FIGURES

Figure B1: Air-Quality Monitor

Sources: https://new.qq.com/rain/a/20170124015370

Figure B2: Map of Monitor Location

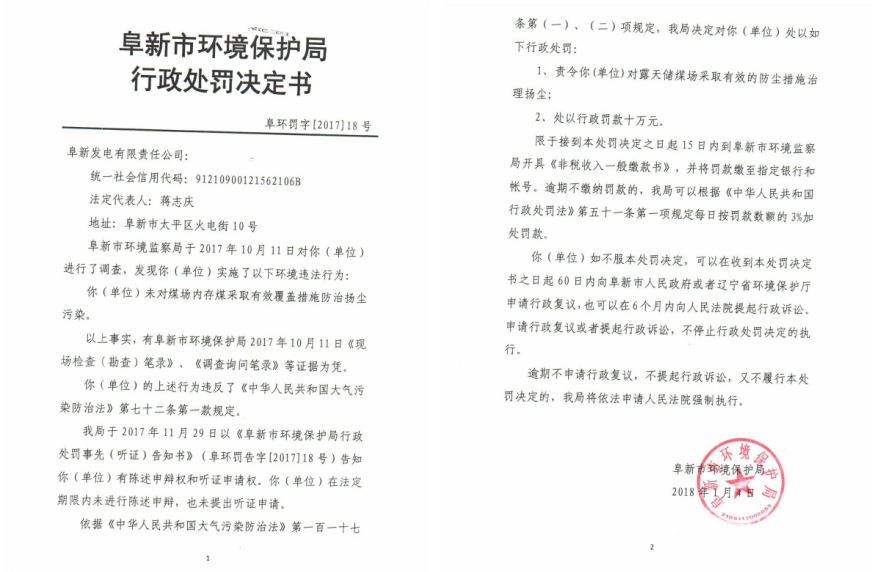


Notes: This figure shows the location of monitors used in the analysis. The main analysis uses monitors installed in 2015 (shown in red on the map). The map also reports the location of monitors installed in the three major urban areas Beijing-Tianjin-Hebei, Pearl River Delta, and Yangtze River Delta (in yellow) as well as the location of monitors in environmental role model cities (in blue).

Figure B3: Aerosol Optical Depth over Time

Notes: This figure shows the change in average Aerosol Optical Depth in China from 2010 to 2018.

Figure B4: An Enforcement Issued by Fuxin Government



阜新市环境保护局
行政处罚决定书

阜环罚字[2017]18号

阜新发电有限责任公司：

统一社会信用代码：91210900121562106B

法定代表人：蒋志庆

地址：阜新市太平区火电街10号

阜新市环境监察局于2017年10月11日对你（单位）进行了调查，发现你（单位）实施了以下环境违法行为：

你（单位）未对煤场内存煤采取有效覆盖措施防治扬尘污染。

以上事实，有阜新市环境保护局2017年10月11日《现场检查（勘查）笔录》、《调查询问笔录》等证据为凭。

你（单位）的上述行为违反了《中华人民共和国大气污染防治法》第七十二条第一款规定。

我局于2017年11月29日以《阜新市环境保护局行政处罚事先（听证）告知书》（阜环罚告字[2017]18号）告知你（单位）有陈述申辩权和听证申请权。你（单位）在法定期限内未进行陈述申辩，也未提出听证申请。

依据《中华人民共和国大气污染防治法》第一百一十七

1

条第（一）、（二）项规定，我局决定对你（单位）处以如下行政处罚：

1、责令你(单位)对露天储煤场采取有效的防尘措施治理扬尘；

2、处以行政罚款十万元。

限于接到本处罚决定之日起15日内到阜新市环境监察局开具《非税收入一般缴款书》，并将罚款缴至指定银行和帐号。逾期不缴纳罚款的，我局可以根据《中华人民共和国行政处罚法》第五十一条第一项规定每日按罚款数额的3%加处罚款。

你（单位）如不服本处罚决定，可以在收到本处罚决定书之日起60日内向阜新市人民政府或者辽宁省环境保护厅申请行政复议，也可以在6个月内向人民法院提起行政诉讼。申请行政复议或者提起行政诉讼，不停止行政处罚决定的执行。

逾期不申请行政复议，不提起行政诉讼，又不履行本处罚决定的，我局将依法申请人民法院强制执行。

阜新市环境保护局

2018年1月4日

2

The decision on administrative penalties from Environmental Protection Agency in Fuxin City

[2017] No. 18

To
Fuxin Electricity Company Limited
Social credit code: 91210900121562106B
Legal representative: Zhiqing Jiang
Address: Huodian Street No. 10, Taiping district, Fuxin city

The Fuxin Environmental Monitoring Bureau investigated you (Fuxin Electricity Company Limited) on the 11th of Oct. in 2017, and found below violations:

You (Fuxin Electricity Company Limited) didn't take effective measure to prevent dust pollution.

Above facts can be verified and checked by the evidences such as site survey record and inquiry record made by Environmental Protection Agency of Fuxin City on the 11th of Oct. in 2017.

Above facts violated the first paragraph of Article 72 of the Law of the People's Republic of China on Prevention and Control of Air Pollution.

We notified you about your right to state, defend and apply for hearing by sending you "The Prior Notice of Administrative Penalties from Environmental Protection Agency in Fuxin City" ([2017] No. 18) on the 29th of Nov. in 2017. You didn't provide any defense and application for hearing within legal period.

According to Regulations (1) and (2) of Article 117 of the Law of the People's Republic of China on the Prevention and Control of Air Pollution, we decided to impose below administrative penalties on you:

1. Order you to take effective measures to prevent dust pollution in open-pit coal storage yard;
2. Administrative fine up to 100,000 yuan.

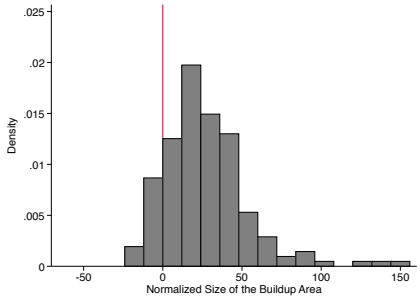
You must present yourself at the Fuxin Environmental Monitoring Bureau to receive "General Non-Tax Income Payments" and pay the fine to the designated bank and account number within 15 days from the date of receipt of this penalty decision. If the fine is not paid within the time limit, the Office may impose an additional fine of 3% of the original fine amount on a daily basis in accordance with the first paragraph of Article 51 of the Administrative Punishment Law of the People's Republic of China.

If you refuses to accept this penalty decision, you may apply to the Fuxin Municipal People's Government or the Liaoning Provincial Environmental Protection Department for administrative reconsideration within 60 days from the date of receipt of this penalty decision. You may also file an administrative lawsuit with the People's Court within 6 months. Applying for administrative reconsideration or filing an administrative lawsuit does not stop the execution of the administrative penalty decision.

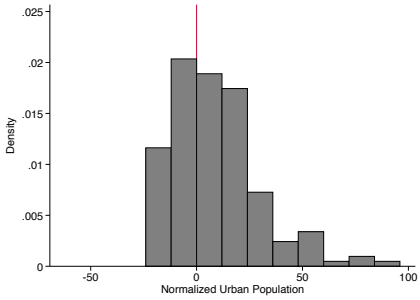
If you do not apply for administrative reconsideration within the time limit, do not file an administrative lawsuit, and fail to perform the decision on this penalty, the bureau will apply to the people's court for compulsory execution according to law.

The Environmental Protection Agency in Fuxin City
4th of Jan, 2018

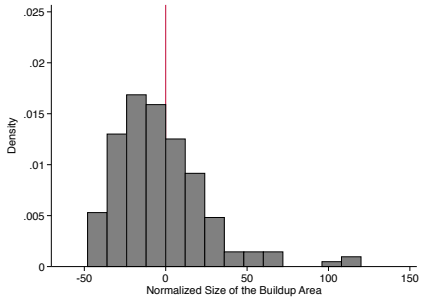
Figure B5: Histogram of Running Variables



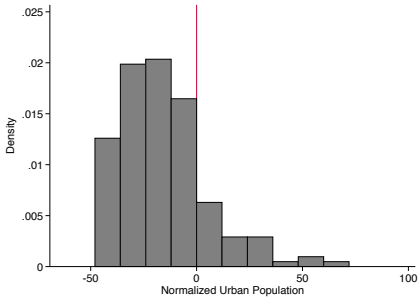
(a) Distribution of the Size of the built-up area in Cutoff 1



(b) Distribution of the Urban Population in Cutoff 1



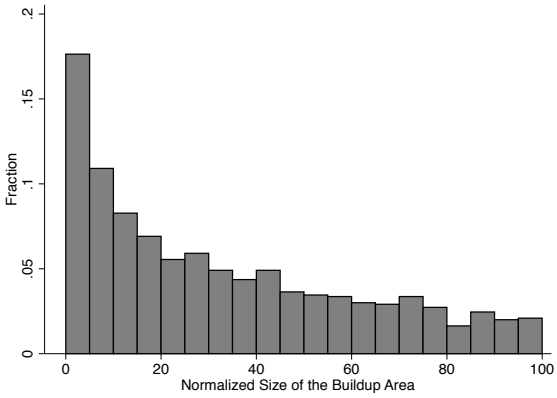
(c) Distribution of the Size of the built-up area in Cutoff 1



(d) Distribution of the Urban Population in Cutoff 2

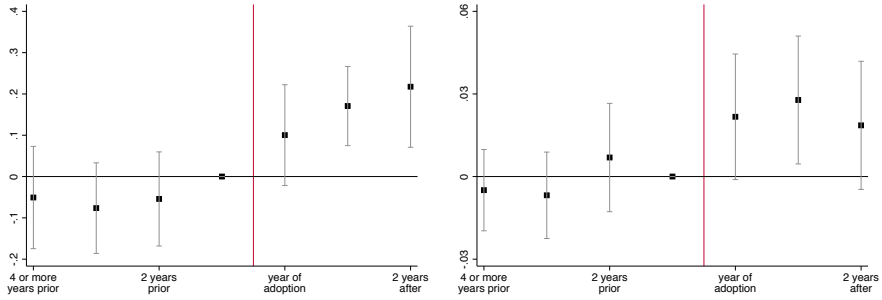
Notes: The figures provide histograms of the urban population and the size of the built-up area of our sample over two cutoffs. Please note that the Size of the built-up area and the Urban Population were normalized.

Figure B6: Histogram of the Distance



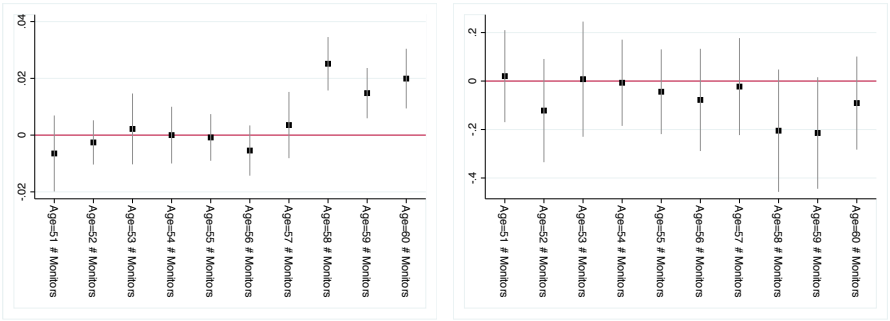
Notes: The figure shows the histogram of the distance between ESR firm and the closest Monitor. We restrict the sample to firms that locate within 100 kilometers.

Figure B7: Event Study



Notes: The figures provide graphic analyses of the effects of monitoring on the enforcement using DiD specification. The left figure display the event study for the estimate of Column 1 in Table 4.7. The right figure display the event study for the estimate of Column 1 in Table 4.8. The figure plots the coefficients of the event study. Capped spikes represent 95 percent confidence intervals. See the text for further details.

Figure B8: Promotion Incentive



(a) AOD (b) Enforcement

Notes: The figures display the effects of an additional monitor on both pollution and enforcement by the age of the mayor in 2018. The effect is normalized to 0 for cities with a mayor who would be 50 years old in 2018. The figure left is the effect on the pollution reduction. The figure right is the effect on the increase of the enforcement.

Figure B9: Manipulation

Sources: http://hsb.hsw.cn/2015-01/20/content_8562907.htm

C. DIFFERENCE-IN-DISCONTINUITIES

We also exploit the longitudinal nature of our data using a "Difference-in-Discontinuities" (or Diff-in-Disc) design Grembi, Nannicini, and Troiano (2016). Several studies in the literature have exploited the longitudinal nature of the data in an RD framework, such as the fixed-effect RD estimator in (Pettersson-Lidbom, 2012), the first-difference RD estimator in (Lemieux and Milligan, 2008), or the dynamic RD design in (Cellini, Ferreira, and Rothstein, 2010). The Diff-in-Disc design essentially combines a difference-in-differences (comparing the air pollution in cities with a different number of monitors, before and after 2015) with a regression discontinuity design (comparing the air pollution of cities just above or below certain cutoffs). To estimate the Diff-in-Disc model, we follow the common practice of using local linear regression. More specifically, we estimate the following equation:

$$AOD_{it} = \delta_0 + \delta_1 D_i + S_i(\gamma_0 + \gamma_1 D_i) + T_t[\alpha_0 + \alpha_1 D_i + S_i(\beta_0 + \beta_1 D_i)] + \xi_{it}, \tag{4.3}$$

where S_i is a dummy for cities above cutoffs, T_t an indicator for the period after 2015, and D_i the normalized running variable. Standard errors are clustered at the city level. Treatment is captured by $T_t \times S_i$ and the coefficient of interest is therefore β_0. This is the Diff-in-Disc estimate and identifies the reduced-form effect of being just above the cutoff. We normalize the estimates to the treatment effect of one additional monitor by dividing β_0 by the first-stage RD estimates. Results of the Diff-in-Disc regressions are shown in Column (4) of the Table 4.6.

BIBLIOGRAPHY

Abadie, Alberto and Guido W Imbens. 2012. "A martingale representation for matching estimators." *Journal of the American Statistical Association* 107 (498):833–843.

Abrevaya, Jason. 2009. "Are There Missing Girls in the United States? Evidence from Birth Data." *American Economic Journal: Applied Economics* 1 (2):1–34.

Acemoglu, Daron, Leopoldo Fergusson, James A Robinson, Dario Romero, and Juan F Vargas. 2018. "The Perils of High-Powered Incentives: Evidence from Colombia's False Positives." *Working paper* .

Adhvaryu, Achyuta, Steven Bednar, Teresa Molina, Quynh Nguyen, and Anant Nyshadham. 2020. "When It Rains It Pours: The Long-Run Economic Impacts of Salt Iodization in the United States." *The Review of Economics and Statistics* 102 (2):395–407.

Adhvaryu, Achyuta, Teresa Molina, Anant Nyshadham, and Jorge Tamayo. 2015. "Helping Children Catch Up: Early Life Shocks and the Progresa Experiment." *Working Paper* .

Agostinelli, Francesco and Matthew Wiswall. 2016. "Estimating the Technology of Children's Skill Formation." *Working Paper* .

Aguilar, Arturo and Marta Vicarelli. 2018. "El Nino and Mexican Children: Medium-Term Effects of Early-Life Weather Shocks on Cognitive and Health Outcomes." *Working Paper* .

Almond, Douglas. 2006. "Is the 1918 Influenza Pandemic Over? Long-term Effects of in Utero Influenza Exposure in the Post-1940 U.S. Population." *Journal of Political Economy* 114 (4):672–712.

Almond, Douglas and Janet Currie. 2011. "Killing Me Softly: The fetal origins hypothesis." *Journal of Economic Perspectives* 25 (3):153–72.

Almond, Douglas, Janet Currie, and Valentina Duque. 2018. "Childhood Circumstances and Adult Outcomes: Act II." *Journal of Economic Literature* 56 (4):1360–1446.

Almond, Douglas and Lena Edlund. 2008. "Son-biased Sex Ratios in the 2000 United States Census." *Proceedings of the National Academy of Sciences* 105 (15):5681–5682.

Almond, Douglas, Lena Edlund, Hongbin Li, and Junsen Zhang. 2010. *Long-Term Effects of Early-Life Development: Evidence From the 1959 to 1961 China Famine*. University of Chicago Press.

Almond, Douglas and Bhashkar Mazumder. 2013. "Fetal Origins and Parental Responses." *Annual Review of Economics* 5 (1):37–56.

Almond, Douglas, Bhashkar Mazumder, and Reyn Van Ewijk. 2015. "In utero Ramadan exposure and children's academic performance." *The Economic Journal* 125 (589):1501–1533.

Andrews, Steven Q. 2008. "Inconsistencies in Air Quality Metrics: 'Blue Sky' Days and PM10 Concentrations in Beijing." *Environmental Research Letters* 3 (3):034009.

Angrist, Joshua and Alan Krueger. 1992. "The Effect of Age at School Entry on Educational Attainment: An Application of Instrumental Variables with Moments from Two Samples." *Journal of the American Statistical Association* 87 (418):328–336.

Angrist, Joshua D and Jörn-Steffen Pischke. 2010. "The credibility revolution in empirical economics: How better research design is taking the con out of econometrics." *Journal of economic perspectives* 24 (2):3–30.

Attanasio, Orazio, Costas Meghir, and Emily Nix. 2018. "Human Capital Development and Parental Investment in India." *Working Paper* .

Attané, Isabelle. 2012. "Being a woman in China today: A demography of gender." *China perspectives* 2012 (2012/4):5–15.

Axbard, Zichen, Sebastianand Deng. 2020. "Information, Accountability and Regulatory Enforcement: Evidence from Pollution Monitoring in China." *Working paper* .

Banerjee, Abhijit V, Esther Duflo, and Rachel Glennerster. 2008. "Putting a Band-Aid on a Corpse: Incentives for Nurses in the Indian Public Health Care System." *Journal of the European Economic Association* 6 (2-3):487–500.

Barcellos, Silvia Helena, Leandro S Carvalho, and Adriana Lleras-Muney. 2014. "Child Gender and Parental Investments in India: Are Boys and Girls Treated Differently?" *American Economic Journal: Applied Economics* 6 (1):157–189.

Barker, David J. 1990. "The Fetal and Infant Origins of Adult Disease." *BMJ: British Medical Journal* 301 (6761):1111.

Barker, David JP and Clive Osmond. 1986. "Infant mortality, childhood nutrition, and ischaemic heart disease in England and Wales." *The Lancet* 327 (8489):1077–1081.

Barwick, Panle Jia, Shanjun Li, Deyu Rao, and Nahim Bin Zahur. 2018. "The Morbidity Cost of Air Pollution: Evidence From Consumer Spending in China." *Working paper* .

Baumann, Eugen. 1896. "Ueber das normale Vorkommen von Jod im Thierkörper.(I. Mittheilung)." *Hoppe-Seyler's Zeitschrift für physiologische Chemie* 21 (4):319–330.

Besley, Timothy and Robin Burgess. 2002. "The Political Economy of Government Responsiveness: Theory and Evidence from India*." *The Quarterly Journal of Economics* 117 (4):1415–1451.

Bharadwaj, Prashant and Leah K Lakdawala. 2013. "Discrimination Begins in the Womb: Evidence of Sex-Selective Prenatal Investments." *Journal of Human Resources* 48 (1):71–113.

Bharadwaj, Prashant, Petter Lundborg, and Dan-Olof Rooth. 2017. "Birth Weight in the Long Run." *Journal of Human Resources* :0715–7235.

Bleakley, Hoyt. 2007. "Disease and Development: Evidence from Hookworm Eradication in the American South." *The Quarterly Journal of Economics* 122 (1):73–117.

———. 2010a. "Health, Human Capital, and Development." *Annual Review of Economics* 2 (1):283–310.

———. 2010b. "Malaria Eradication in the Americas: A Retrospective Analysis of Childhood Exposure." *American Economic Journal: Applied Economics* 2 (2):1–45.

Bombardini, Matilde and Bingjing Li. 2016. "Trade, Pollution and Mortality in China." *Working paper* .

Brown, D, A Kowalski, and I Lurie. 2018. "Long-Term Impacts of Childhood Medicaid Expansions on Outcomes in Adulthood." *Working Paper* .

Buja, Andreas, Lawrence Brown, Richard Berk, Edward George, Emil Pitkin, Mikhail Traskin, Kai Zhang, and Linda Zhao. 2019. "Models as Approximations I: Consequences Illustrated with Linear Regression." *Statist. Sci.* 34 (4):523–544.

Calonico, Sebastian, Matias D Cattaneo, and Rocio Titiunik. 2014. "Robust Nonparametric Confidence Intervals for Regression-Discontinuity Designs." *Econometrica* 82 (6):2295–2326.

Cameron, A Colin, Jonah B Gelbach, and Douglas L Miller. 2008. "Bootstrap-Based Improvements for Inference with Clustered Errors." *The Review of Economics and Statistics* 90 (3):414–427.

Cao, Jiarui, Yiqing Xu, and Chuanchuan Zhang. 2020. "Clans and Calamity: How Social Capital Saves Lives during China's Great Famine." *Working paper* .

Cao, Xue-Yi, Xin-Min Jiang, Zhi-Hong Dou, Murdon Abdul Rakeman, Ming-Li Zhang, Karen O'donnell, Tai Ma, Kareem Amette, Nancy DeLong, and G Robert DeLong. 1994. "Timing of Vulnerability of the Brain to Iodine Deficiency in Endemic Cretinism." *New England Journal of Medicine* 331 (26):1739–1744.

Cellini, Stephanie Riegg, Fernando Ferreira, and Jesse Rothstein. 2010. "The Value of School Facility Investments: Evidence from a Dynamic Regression Discontinuity Design." *The Quarterly Journal of Economics* 125 (1):215–261.

Chen, Jidong, Jennifer Pan, and Yiqing Xu. 2016. "Sources of Authoritarian Responsiveness: A Field Experiment in China." *American Journal of Political Science* 60 (2):383–400.

Chen, Junshi and Huiyun Wu. 1998. "Fortification of Salt with Iodine." *Food and Nutrition Bulletin* 19.

Chen, Yuyu, Ginger Zhe Jin, Naresh Kumar, and Guang Shi. 2013. "The Promise of Beijing: evaluating the impact of the 2008 Olympic Games on air quality." *Journal of Environmental Economics and Management* 66 (3):424–443.

Chen, Yuyu and Li-An Zhou. 2007. "The Long-Term Health and Economic Consequences of the 1959–1961 Famine in China." *Journal of Health Economics* 26 (4):659–681.

Chen, Yvonne Jie, Pei Li, and Yi Lu. 2018. "Career Concerns and Multitasking Local Bureaucrats: Evidence of a Target-based Performance Evaluation System in China." *Journal of Development Economics* 133:84–101.

Chetty, Raj, John N Friedman, Nathaniel Hilger, Emmanuel Saez, Diane Whitmore Schanzenbach, and Danny Yagan. 2011. "How Does Your Kindergarten Classroom Affect Your Earnings? Evidence from Project STAR." *The Quarterly Journal of Economics* 126 (4):1593–1660.

Conley, Timothy G, Christian B Hansen, and Peter E Rossi. 2012. "Plausibly exogenous." *Review of Economics and Statistics* 94 (1):260–272.

Cunha, Flavio, James J Heckman, Lance Lochner, and Dimitriy V Masterov. 2006. "Interpreting the Evidence on Life Cycle Skill Formation." *Handbook of the Economics of Education* 1:697–812.

Cunha, Flavio, James J Heckman, and Susanne M Schennach. 2010. "Estimating the Technology of Cognitive and Noncognitive Skill Formation." *Econometrica* 78 (3):883–931.

Dahl, Gordon B, Andreas Kotsadam, and Dan-Olof Rooth. 2017. "Does Integration Change Gender Attitudes?" *Working Paper* .

Deng, Zichen and Maarten Lindeboom. 2019. "A Bit of Salt, A Trace of Life: Gender Norms and The Impact of a Salt Iodization Program on Human Capital Formation of School Aged Children." *Working paper* .

———. 2020. "Early-life Famine Exposure, Hunger Recall and Later-life Health." *Working paper* .

Dhaliwal, Iqbal and Rema Hanna. 2017. "The Devil is in the Details: The Successes and Limitations of Bureaucratic Reform in India." *Journal of Development Economics* 124:1 – 21.

Dhar, Diva, Tarun Jain, and Seema Jayachandran. 2018. "Reshaping Adolescents' Gender Attitudes: Evidence from a School-Based Experiment in India." *Working Paper* .

Dinkelman, Taryn. 2017. "Long-run Health Repercussions of Drought Shocks: Evidence from South African Homelands." *The Economic Journal* 127 (604):1906–1939.

Dossi, Gaia, David N Figlio, Paola Giuliano, and Paola Sapienza. 2019. "Born in the Family: Preferences for Boys and the Gender Gap in Math." *Working Paper* .

Drange, Nina and Tarjei Havnes. 2019. "Early Childcare and Cognitive Development: Evidence from an Assignment Lottery." *Journal of Labor Economics* 37 (2):581–620.

Duflo, Esther, Michael Greenstone, Rohini Pande, and Nicholas Ryan. 2013. "Truth-Telling by Third-Party Auditors and the Response of Polluting Firms: Experimental Evidence from India." *The Quarterly Journal of Economics* 128 (4):1499–1545.

———. 2018. "The Value of Regulatory Discretion: Estimates From Environmental Inspections in India." *Econometrica* 86 (6):2123–2160.

Duflo, Esther, Rema Hanna, and Stephen P. Ryan. 2012. "Incentives Work: Getting Teachers to Come to School." *American Economic Review* 102 (4):1241–78.

Duque, V, MF Rosales, and F Sanchez. 2018. "How Do Early-Life Shocks Interact with Subsequent Human-Capital Investments? Evidence from Administrative Data." *Working Paper* .

Ebenstein, Avraham, Maoyong Fan, Michael Greenstone, Guojun He, and Maigeng Zhou. 2017. "New Evidence on the Impact of Sustained Exposure to Air Pollution on Life Expectancy from China's Huai River Policy." *Proceedings of the National Academy of Sciences* 114 (39):10384–10389.

Edlund, Lena, Hongbin Li, Junjian Yi, and Junsen Zhang. 2013. "Sex Ratios and Crime: Evidence from China." *Review of Economics and Statistics* 95 (5):1520–1534.

Felfe, Christina and Rafael Lalive. 2018. "Does Early Child Care Affect Children's Development?" *Journal of Public Economics* 159:33–53.

Feyrer, James, Dimitra Politi, and David N Weil. 2017. "The Cognitive Effects of Micronutrient Deficiency: Evidence from Salt Iodization in the United States." *Journal of the European Economic Association* 15 (2):355–387.

Field, Erica, Omar Robles, and Maximo Torero. 2009. "Iodine Deficiency and Schooling Attainment in Tanzania." *American Economic Journal: Applied Economics* 1 (4):140–169.

Figlio, David, Jonathan Guryan, Krzysztof Karbownik, and Jeffrey Roth. 2014. "The Effects of Poor Neonatal Health on Children's Cognitive Development." *American Economic Review* 104 (12):3921–3955.

Fisman, Raymond and Yongxiang Wang. 2017. "The Distortionary Effects of Incentives in Government: Evidence from China's D̈eath CeilingP̈rogram." *American Economic Journal: Applied Economics* 9 (2):202–18.

Ghanem, Dalia and Junjie Zhang. 2014. "'Effortless Perfection:'Do Chinese cities manipulate air pollution data?" *Journal of Environmental Economics and Management* 68 (2):203–225.

Greenstone, Michael, He Guojun, Jia Ruixue, and Liu Tong. 2019. "Can Technology Solve the Principal-Agent Problem? Evidence from Pollution Monitoring in China." *Working paper* .

Greenstone, Michael and Rema Hanna. 2014. "Environmental Regulations, Air and Water Pollution, and Infant Mortality in India." *American Economic Review* 104 (10):3038–72.

Greenstone, Michael and Patrick Schwarz. 2018. "Is China Winning its War on Pollution?" Tech. rep., Energy Policy Institute at the University of Chicago (EPIC).

Grembi, Veronica, Tommaso Nannicini, and Ugo Troiano. 2016. "Do Fiscal Rules Matter?" *American Economic Journal: Applied Economics* 8 (3):1–30.

Grönqvist, Hans, J Peter Nilsson, Per-Olof Robling et al. 2018. "Early Lead Exposure and Outcomes in Adulthood." *Working Paper* .

Gunnsteinsson, Snaebjorn, Achyuta Adhvaryu, Parul Christian, Alain Labrique, Jonathan Sugimoto, Abu Ahmed Shamim, and Keith P West Jr. 2018. "Protecting Infants from Natural Disasters: The Case of Vitamin A Supplementation and a Tornado in Bangladesh." *Working Paper* .

Gupta, Pawan, Sundar A Christopher, Jun Wang, Robert Gehrig, YC Lee, and Naresh Kumar. 2006. "Satellite Remote Sensing of Particulate Matter and Air Quality Sssessment Over Global Cities." *Atmospheric Environment* 40 (30):5880–5892.

He, Guojun, Shaoda Wang, and Bing Zhang. 2020. "Watering Down Environmental Regulation in China." *The Quarterly Journal of Economics* 135 (4):2135–2185.

Heckman, James J. 2007. "The Economics, Technology, and Neuroscience of Human Capability Formation." *Proceedings of the National Academy of Sciences* 104 (33):13250–13255.

Heffelfinger, Amy K. and John W. Newcomer. 2001. "Glucocorticoid effects on memory function over the human life span." *Development and Psychopathology* 13 (3):491–513.

Henderson, J. Vernon, Adam Storeygard, and David N. Weil. 2012. "Measuring Economic Growth from Outer Space." *American Economic Review* 102 (2):994–1028.

Ho, Daniel E, Kosuke Imai, Gary King, and Elizabeth A Stuart. 2007. "Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference." *Political analysis* 15 (3):199–236.

Hoynes, Hilary, Diane Whitmore Schanzenbach, and Douglas Almond. 2016. "Long-run impacts of childhood access to the safety net." *American Economic Review* 106 (4):903–34.

Huang, Cheng, Zhu Li, Meng Wang, and Reynaldo Martorell. 2010. "Early life exposure to the 1959–1961 Chinese famine has long-term health consequences." *The Journal of nutrition* 140 (10):1874–1878.

Ichino, Andrea, Margherita Fort, and Giulio Zanella. 2019. "Cognitive and Non-Cognitive Costs of Daycare 0-2 for Children in Advantaged Families." *Working Paper* .

Inoue, Atsushi and Gary Solon. 2010. "Two-sample Instrumental Variables Estimators." *The Review of Economics and Statistics* 92 (3):557–561.

Jayachandran, Seema. 2015. "The Roots of Gender Inequality in Developing Countries." *Annual Review of Economics* 7 (1):63–88.

Jayachandran, Seema and Ilyana Kuziemko. 2011. "Why Do Mothers Breastfeed Girls Less Than Boys? Evidence and Implications for Child Health in India." *The Quarterly Journal of Economics* 126 (3):1485–1538.

Jia, Ruixue. 2017. "Pollution for Promotion." *Working paper* .

Jiang, Junyan. 2018. "Making Bureaucracy Work: Patronage Networks, Performance Incentives, and Economic Development in China." *American Journal of Political Science* 62 (4):982–999.

Jooste, Pieter L, Michael J Weight, and Carl J Lombard. 2000. "Short-term Effectiveness of Mandatory Iodization of Table Salt, at an Elevated Iodine Concentration, on the Iodine and Goiter Status of Schoolchildren with Endemic Goiter." *The American Journal of Clinical Nutrition* 71 (1):75–80.

Kahn, Matthew E., Pei Li, and Daxuan Zhao. 2015. "Water Pollution Progress at Borders: The Role of Changes in China's Political Promotion Incentives." *American Economic Journal: Economic Policy* 7 (4):223–42.

Kim, Seonghoon, Belton Fleisher, and Jessica Ya Sun. 2017. "The Long-term health effects of fetal malnutrition: Evidence from the 1959–1961 China great leap forward famine." *Health economics* 26 (10):1264–1277.

Kling, Jeffrey R, Jeffrey B Liebman, and Lawrence F Katz. 2007. "Experimental Analysis of Neighborhood Effects." *Econometrica* 75 (1):83–119.

Kosack, Stephen and Archon Fung. 2014. "Does Transparency Improve Governance?" *Annual Review of Political Science* 17 (1):65–87.

Kung, James Kai-sing and Justin Yifu Lin. 2003. "The causes of China's great leap famine, 1959–1961." *Economic Development and Cultural Change* 52 (1):51–73.

Lemieux, Thomas and Kevin Milligan. 2008. "Incentive Effects of Social Assistance: A Regression Discontinuity Approach." *Journal of Econometrics* 142 (2):807 – 828.

Li, Chihua and LH Lumey. 2017. "Exposure to the Chinese famine of 1959-61 in early life and long-term health conditions: a systematic review and meta-analysis." *International journal of epidemiology* 46 (4):1157–1170.

Li, Yanping, Vincent W Jaddoe, Lu Qi, Yuna He, Jianqiang Lai, Jiansheng Wang, Jian Zhang, Yisong Hu, Eric L Ding, Xiaoguang Yang et al. 2011. "Exposure to the Chinese famine in early life and the risk of hypertension in adulthood." *Journal of Hypertension* 29 (6):1085–1092.

Lin, Justin Yifu and Dennis Tao Yang. 2000. "Food Availability, Entitlements and the Chinese Famine of 1959–61." *The Economic Journal* 110 (460):136–158.

Liu, Yu. 2017. "Informal Taxation and Firm Performance: Evidence from China." *Working paper* .

Lumey, Lambert H, Aryeh D Stein, and Ezra Susser. 2011. "Prenatal Famine and Adult Health." *Annual Review of Public Health* 32 (1):237–262.

Maccini, Sharon and Dean Yang. 2009. "Under the weather: Health, schooling, and economic consequences of early-life rainfall." *American Economic Review* 99 (3):1006–26.

Malamud, Ofer, Cristian Pop-Eleches, and Miguel Urquiola. 2016. "Interactions Between Family and School Environments: Evidence on Dynamic Complementarities?" *Working Paper* .

Meng, Tianguang, Jennifer Pan, and Ping Yang. 2017. "Conditional Receptivity to Citizen Participation: Evidence from a Survey Experiment in China." *Comparative Political Studies* 50 (4):399–433.

Meng, Xin and Nancy Qian. 2009. "The Long Term Consequences of Famine on Survivors: Evidence from a Unique Natural Experiment using China's Great Famine." *Working paper* .

Meng, Xin, Nancy Qian, and Pierre Yared. 2015. "The Institutional Causes of China's Great Famine, 1959–1961." *The Review of Economic Studies* 82 (4):1568–1611.

Michalopoulos, Stelios and Elias Papaioannou. 2018. "Spatial Patterns of Development: A Meso Approach." *Annual Review of Economics* 10 (1):383–410.

Neidell, Matthew and Janet Currie. 2005. "Air Pollution and Infant Health: What Can We Learn from California's Recent Experience?*." *The Quarterly Journal of Economics* 120 (3):1003–1030.

Niemesh, Gregory T. 2015. "Ironing Out Deficiencies: Evidence from the United States on the Economic Effects of Iron Deficiency." *Journal of Human Resources* 50 (4):910–958.

Ohlin, Bertil. 1938. "Economic Progress in Sweden." *The Annals of the American Academy of Political and Social Science* .

Oliva, Paulina. 2015. "Environmental Regulations and Corruption: Automobile Emissions in Mexico City." *Journal of Political Economy* 123 (3):686–724.

Olken, Benjaminŏ0a0A. 2007. "Monitoring Corruption: Evidence from a Field Experiment in Indonesia." *Journal of Political Economy* 115 (2):200–249.

O'malley, Patrick M., Jerald G. Bachman, and Lloyd D. Johnston. 1983. "Reliability and Consistency in Self-Reports of Drug Use." *International Journal of the Addictions* 18 (6):805–824. PMID: 6605313.

Oster, Emily. 2009. "Does Increased Access Increase Equality? Gender and Child Health Investments in India." *Journal of Development Economics* 89 (1):62–76.

Pardede, Lucia VH, Widanto Hardjowasito, Rainer Gross, Drupadi HS Dillon, Ongko S Totoprajogo, Mardhani Yosoprawoto, Lies Waskito, and Juliawati Untoro. 1998. "Urinary Iodine Excretion is the Most Appropriate Outcome Indicator for Iodine Deficiency at Field Conditions at District Level." *The Journal of Nutrition* 128 (7):1122–1126.

Pathania, Vikram. 2007. "The Long Run Impact of Drought at Birth on Height of Women in Rural India." *Working Paper* .

Pettersson-Lidbom, Per. 2012. "Does the Size of the Legislature Affect the Size of Government? Evidence from Two Natural Experiments." *Journal of Public Economics* 96 (3):269 – 278.

Przeworski, Adam. 1986. *Capitalism and Social Democracy*. Cambridge University Press.

Qin, Yu and Hongjia Zhu. 2018. "Run Away? Air Pollution and Emigration Interests in China." *Journal of Population Economics* 31 (1):235–266.

Ravallion, Martin. 1997. "Famines and Economics." *Journal of Economic Literature* 35 (3):1205–1242.

Reinikka, Ritva and Jakob Svensson. 2005. "Fighting Corruption to Improve Schooling: Evidence from a Newspaper Campaign in Uganda." *Journal of the European Economic Association* 3 (2Ď080Ď00903):259–267.

———. 2011. "The Power of Information in Public Services: Evidence from Education in Uganda." *Journal of Public Economics* 95 (7-8):956–966.

Ridder, Geert and Robert Moffitt. 2007. "The Econometrics of Data Combination." *Handbook of Econometrics* 6:5469–5547.

Roseboom. 2010. *Baby's van de Hongerwinter. De onvermoede erfenis van ondervoeding*. Augustus.

Roseboom, Tessa, Susanne de Rooij, and Rebecca Painter. 2006. "The Dutch famine and its long-term consequences for adult health." *Early human development* 82 (8):485–491.

Rossin-Slater, Maya and Miriam Wüst. 2018. "What is the Added Value of Preschool? Long-Term Impacts and Interactions with an Infant Health Intervention." *Working Paper* .

Schlenker, Wolfram and W Reed Walker. 2015. "Airports, Air Pollution, and Contemporaneous health." *The Review of Economic Studies* 83 (2):768–809.

Scholte, Robert S, Gerard J Van den Berg, and Maarten Lindeboom. 2015. "Long-run Effects of Gestation During the Dutch Hunger Winter Famine on Labor Market and Hospitalization Outcomes." *Journal of Health Economics* 39:17–30.

Selten, Jean-Paul, Yolanda van der Graaf, Rozemarijn van Duursen, Christien C Gispen de Wied, and RenÃĺ S Kahn. 1999. "Psychotic illness after prenatal exposure to the 1953 Dutch Flood Disaster." *Schizophrenia Research* 35 (3):243–245.

Sen, Amartya. 1981. *Poverty and Famines: An Essay on Entitlement and Deprivation*. Oxford University Press.

Shah, Manisha and Bryce Millett Steinberg. 2017. "Drought of Opportunities: Contemporaneous and Long-Term Impacts of Rainfall Shocks on Human Capital." *Journal of Political Economy* 125 (2):527–561.

Shu, Heng and Zhiqiang Tan. 2020. "Improved methods for moment restriction models with data combination and an application to two-sample instrumental variable estimation." *Canadian Journal of Statistics* .

Simoons, Frederick J. 1990. *Food in China: A Cultural and Historical Inquiry*. CRC Press.

Snyder, James0̆0a0M. and David Strömberg. 2010. "Press Coverage and Political Accountability." *Journal of Political Economy* 118 (2):355–408.

Staiger, Douglas and Harold James. 1997. "Instrumental Variables with Weak Instruments." *Econometrica* 65 (3):557–586.

Stein, Aryeh D, Henry S Kahn, Andrew Rundle, Patricia A Zybert, Karin van der Pal-de Bruin, and LH Lumey. 2007. "Anthropometric Measures in Middle Age after exposure to famine during gestation: evidence from the Dutch famine." *The American Journal of Clinical Nutrition* 85 (3):869–876.

Stein, Zena, Mervyn Susser, Gerhart Saenger, and Francis Marolla. 1972. "Nutrition and Mental Performance." *Science* 178 (4062):708–713.

Sun, Dianjun. 2018. *Endemic Disease in China*. Springer.

Tan, Chih Ming, Tan Zhibo, and Xiaobo Zhang. 2015. "Sins of the Fathers: The Intergenerational Legacy of the 1959-61 Great Chinese Famine on Children's Cognitive Development." *Working paper* .

Tanaka, Shinsuke. 2015. "Environmental Regulations on Air Pollution in China and Their Impact on Infant Mortality." *Journal of Health Economics* 42:90 – 103.

UN. 2019. "Environmental Rule of Law: First Global Report." Tech. rep., United Nations.

Van den Berg, Gerard J and Maarten Lindeboom. 2018. "Famines, Hunger, and Later-Life Health." *The Oxford Research Encyclopedia of Economics and Finance* .

Van den Berg, Gerard J, Maarten Lindeboom, and France Portrait. 2006. "Economic Conditions Early in Life and Individual Mortality." *The American Economic Review* 96 (1):290–302.

Van den Berg, Gerard J, Pia R Pinger, and Johannes Schoch. 2016. "Instrumental Variable Estimation of the Causal Effect of Hunger Early in Life on Health Later in Life." *The Economic Journal* 126 (591):465–506.

van Kippersluis, Hans and Cornelius A. Rietveld. 2018. "Beyond plausibly exogenous." *The Econometrics Journal* 21 (3):316–331.

Vansteelandt, Stijn and Vanessa Didelez. 2018. "Improving the robustness and efficiency of covariate-adjusted linear instrumental variable estimators." *Scandinavian Journal of Statistics* 45 (4):941–961.

Wang, Jun and Sundar A Christopher. 2003. "Intercomparison Between Satellite-Derived Aerosol Optical Thickness and PM2.5 mass: Implications for air quality studies." *Geophysical Research Letters* 30 (21).

Wang, Le. 2013. "Estimating Returns to Education When the IV Sample is Selective." *Labour Economics* 21:74 – 85.

Woo, Wing Thye, Shi Li, Ximing Yue, Harry Xiaoying Wu, and Xinpeng Xu. 2004. "The Poverty Challenge for China in the New Millennium." *The Millennium Development Goals Project of the United Nations* .

Xi, Tianyang, Yang Yao, and Muyang Zhang. 2018. "Capability and Opportunism: Evidence from City Officials in China." *Journal of Comparative Economics* forthcoming.

Xu, Hongwei, Lydia Li, Zhenmei Zhang, and Jinyu Liu. 2016. "Is natural experiment a cure? Re-examining the long-term health effects of China's 1959–1961 famine." *Social Science & Medicine* 148:110–122.

Yang, Dali L. 1998. *Calamity and reform in China: State, rural society, and institutional change since the Great Leap Famine*. Stanford University Press.

Yao, Shujie. 1999. "A Note on the Causal Factors of China's Famine in 1959–1961." *Journal of Political Economy* 107 (6):1365–1369.

Yi, Junjian, James J Heckman, Junsen Zhang, and Gabriella Conti. 2015. "Early Health Shocks, Intra-household Resource Allocation and Child Outcomes." *The Economic Journal* 125 (588):F347–F371.

Zhang, Junsen. 2017. "The Evolution of China's One-Child Policy and Its Effects on Family Outcomes." *Journal of Economic Perspectives* 31 (1):141–60.

Zhao, Qingyuan, Jingshu Wang, Wes Spiller, Jack Bowden, Dylan S Small et al. 2019. "Two-Sample Instrumental Variable Analyses Using Heterogeneous Samples." *Statistical Science* 34 (2):317–333.

Zheng, Siqi and Matthew E Kahn. 2013. "Understanding China's Urban Pollution Dynamics." *Journal of Economic Literature* 51 (3):731–72.

———. 2017. "A New Era of Pollution Progress in Urban China?" *Journal of Economic Perspectives* 31 (1):71–92.

Zimmermann, Michael B. 2008. "Research on Iodine Deficiency and Goiter in the 19th and Early 20th Centuries." *Journal of Nutrition* 138 (11):2060–2063.

———. 2011. "The Role of Iodine in Human Growth and Development." In *Seminars in Cell & Developmental Biology*, vol. 22. Elsevier, 645–652.

Zimmermann, Michael B, Sonja Y Hess, Pierre Adou, Toni Toresanni, Rita Wegmüller, and Richard F Hurrell. 2003. "Thyroid Size and Goiter Prevalence after Introduction of Iodized Salt: A 5-y Prospective Study in Schoolchildren in Côte d'Ivoire." *The American Journal of Clinical Nutrition* 77 (3):663–667.

SUMMARY

This thesis consists of three empirical studies that center around development, environmental, and health economics. This chapter summarizes the main findings and conclusions from the previous chapters.

Chapter 2 leverages newly collected individual-level hunger recall information of the China Family Panel Survey to estimate the causal effect of undernourishment on later-life health. We develop a Two-Sample Instrumental Variable (TSIV) estimator that can deal with heterogeneous samples. The new estimator includes a first step which pre-processes the data by methods such as the nearest-neighbor matching. The matching, which homogenizes covariates' distributions in different samples, decreases the parametric choice's dependence in the second-step regression. Using the new estimator, we find a nonlinear relationship between the widely used indicator of famine intensity and personal exposure (measured by hunger recall among famine survivors). The nonlinearity in famine exposure may explain the variation in the famine's effect on later life health found in previous studies, as studies explore different parts of this non-linear relation. We furthermore find that exposure to famine induced hunger early in life leads to worse health among females fifty years later. This effect is more significant than the reduced-form effect of previous studies. For males, we find no impact. This chapter belongs to the mounting concurrent evidence that malnutrition around birth affects a broad range of late-life health outcomes. However, previous studies are significantly constrained as the information on individual exposure is not available. In our study, the information on personal exposure helps in the justification of instruments used in this literature, as well as providing insight into the proper specification of the reduced-form regressions used in the extensive famine literature. More importantly, the lack of information on personal exposure is not unique in the famine literature. The method developed in this paper can be used in other similar studies, primarily when key variables for identification are constructed from historical information. The strategy complements the efforts in collecting data on exogenous variation.

Chapter 3 examines the effects of a massive salt iodization program on the human capital formation of school-aged children in China. To identify the salt iodizing policy's long-term benefits, we use the national salt iodizing program as a quasi-experiment and

exploit geographic variation in goiter prevalence before the intervention. So, we essentially compare improvements in math and vocabulary ability and educational attainment and years of schooling of cohorts conceived before and after the salt iodization in provinces with varying pre-intervention goiter prevalence. Our difference-in-differences estimates show that the salt iodization policy has strong and significant effects on cognition for girls. We find robust positive effects of the program for girls. A one standard deviation decrease (12%) in the pre-intervention goiter rate is associated with math and vocabulary scores increasing by roughly 15%. We also see substantial increases in the educational attainment of females. Using a simple back of the envelope calculation, we infer that this translates to income increases of about 6%. Yet, we do not find any effects for boys. We show in a simple model of parental investment that gender preferences can explain our findings. We consider the role of gender preferences and how this may affect large-scale public programs' effectiveness. Gender preferences may also explain gender differences in the existing empirical literature on the long-run effects of adverse conditions early in life. Analyses exploiting within the province, village-level variation in gender attitudes find that the gains in cognition are most significant for girls born in regions with the strongest son preferences. The policy's heterogeneous impact confirms the importance of parental gender preferences. Consequently, large scale programs can have positive (and possibly) unintended effects on gender equality in societies with son preference.

Chapter 4 turns the research focus to the implementation of environmental regulations in developing countries. Despite ambitious environmental laws in many countries worldwide, the enforcement of these regulations is often weak. Holding government officials accountable for this lack of enforcement is, in turn, often marred by inadequate information about environmental quality. In this chapter, we study whether better environmental monitoring can solve this issue and improve the policy's effectiveness. We focus on air pollution in China and investigate the impact of a nationwide monitoring program. Using identification strategies that exploit strict assignment criteria set up by the central government, we show that an additional air pollution monitor reduced satellite-based pollution measures by 2-3% and increased enforcement of air-pollution regulations by 20%. To clarify how this pollution reduction came about, we construct a novel data set of city governments' air pollution enforcement activities from 2010 to 2017. We geo-reference enforcement activities and show that the increase in enforcement is driven by firms located close to those monitors, whereas firms faraway face fewer enforcement changes. These effects are caused by local officials that face strong incentives to reduce pollution and are stronger when there is limited scope for data manipu-

lation – suggesting that better quality information can strengthen the accountability of implementing officials and improve policy outcomes.

SAMENVATTING

Dit proefschrift bestaat uit drie empirische studies rondom ontwikkelings-, milieu- en gezondheidseconomie. Dit hoofdstuk vat de belangrijkste bevindingen en conclusies uit de voorgaande hoofdstukken samen.

Hoofdstuk 2 maakt gebruike van nieuw verzamelde gegevens op individueel niveau over herinneringen aan honger van de China Family Panel Survey om de causale effecten van ondervoeding op gezondheid in het latere leven te schatten. We ontwikkelen een Two-Sample Instrumental Variable (TSIV) schatter die kan omgaan met heterogene steekproeven. Deze nieuwe schatter bevat een eerste stap de de gegevens voorbewerkt met methodes zoals de nearest neighbor matching. De "matching" methode, die de verdelingen van covariaten in verschillende steekproeven homogeniseert, vermindert de afhankelijkheid van de parametrische keuze in de regressie in de tweede stap. Met de nieuwe schatter vinden we een niet-lineaire relatie tussen de de veelgebruikte indicator van hongersnood-intensiteit en persoonlijke blootstelling (gemeten aan de hand van herinneringen aan honger onder overlevenden van de hongersnood). De niet-lineariteit in blootstelling aan hongersnood kan de variatie verklaren in het effect van de hongersnood op gezondheidsuitkomsten op latere leeftijd die in eerdere studies werd gevonden, aangezien studies verschillende delen van deze niet-lineaire relatie onderzoeken. Verder vinden we dat blootstelling op jonge leeftijd aan door hongersnood veroorzaakte honger leidt tot slechtere gezondheid van vrouwen vijftig jaar later. Dit effects is significanter dan het "reduced-form" effect van eerdere onderzoeken. Voor mannen vinden we geen impact. Dit hoofdstuk behoort tot het groeiende bewijs dat ondervoeding rond de geboorte een breed scala aan gezondheidsuitkomsten op latere leeftijd beïnvloedt. Eerdere onderzoeken zijn echter aanzienlijk beperkt omdat informatie over individuele blootstelling niet beschikbaar is in deze onderzoeken. In ons onderzoek helpt de informatie over persoonlijke blootstelling bij het rechtvaardigen van instrumentele variabelen die in deze literatuur worden gebruikt, en geeft het inzicht in de juiste specificatie van de "reduced-form" regressies die in de uitgebreide hongersnood-literatuur worden gebruikt. Nog belangrijker is dat het gebrek aan informatie over persoonlijke blootstelling niet uniek is in de hongersnood-literatuur. De methode ontwikkeld in dit artikel kan in andere vergelijkbare onderzoeken worden gebruikt, voornamelijk wanneer de belangri-

jkste variabelen voor identificatie worden geconstrueerd op basis van historische informatie. De strategie vormt een aanvulling op inspanningen om gegevens over exogene variatie te verzamelen.

Hoofdstuk 3 onderzoekt de effecten van een grootschalig overheidsprogramma waarbij jodium aan zout is toegevoegd (jodiumprogramma) op de ontwikkeling van menselijk kapitaal bij schoolgaande kinderen in China. Om de langetermijnvoordelen van beleid om jodium toe te voegen aan zout te identificeren, gebruiken we het nationale jodiumprogramma als een quasi-experiment en benutten we geografische variatie in de prevalentie van struma vóór de interventie. We vergelijken dus in wezen verbeteringen in de wiskundige en vocabulaire bekwaamheid, opleidingsniveau en het aantal schooljaren van cohorten die zijn verwekt voor en na de invoering van het beleid om jodium toe te voegen aan zout in provincies met een variërende prevalentie van struma vóór de interventie. Onze verschil-in-verschillen schattingen laten zien dat het jodiumbeleid sterke en significante effecten heeft op de cognitie van meisjes. We vinden robuuste positieve effecten van het programma voor meisjes. Een afname van één standaarddeviatie (12%) in het percentage struma bij kinderen vóór de interventie wordt geassocieerd met een stijging van de scores voor wiskunde en woordenschat met ongeveer 15%. We zien ook substantiële stijgingen in het behaalde opleidingsniveau van vrouwen. Met behulp van een eenvoudige back-of-the-envelope berekening, leiden we af dat dit zich vertaalt in inkomensstijgingen van ongeveer 6 %. We vinden echter geen effecten voor jongens. We laten in een eenvoudig model van ouderlijke investeringen zien dat geslachtsvoorkeuren onze bevindingen kunnen verklaren. We nemen de rol van geslachtsvoorkeuren in aanmerking en de wijze waarop dit de effectiviteit van grootschalige publieke programma's kan beïnvloeden. Geslachtsvoorkeuren kunnen ook geslachtsverschillen verklaren in de bestaande empirische literatuur over de langetermijneffecten van ongunstige omstandigheden op jonge leeftijd. Analyses die binnen de provincie gebruik maken van variatie in gender-opvattingen op dorpsniveau benutten, de verbeteringen in cognitie het meest significant zijn voor meisjes die geboren zijn in regio's met de sterkste geslachtsvoorkeuren voor een zoon. De heterogene impact van het beleid bevestigt het belang van geslachtsvoorkeuren van ouders. Zodoende kunnen grootschalige programma's positieve (en mogelijk) onbedoelde effecten hebben op geslachtsgelijkheid in samenlevingen met een voorkeur voor zonen.

Hoofdstuk 4 richt de aandacht van het onderzoek op de implementatie van milieuregelgeving in ontwikkelingslanden. Ondanks ambitieuze milieuwetgeving in veel landen over de hele wereld, is de handhaving van deze regulering vaak zwak. Het verantwoordelijk houden van overheidsfunctionarissen voor dit gebrek aan handhaving wordt

op zijn beurt vaak belemmerd door ontoereikende informatie over de milieukwaliteit. In dit hoofdstuk onderzoeken we of een betere monitoring van het milieu dit probleem kan oplossen en de effectiviteit van het beleid kan verbeteren. We richten ons op luchtverontreiniging in China en onderzoeken de impact van een landelijk monitoringsprogramma. Met behulp van identificatiestrategieën die gebruik maken van strikte toewijzingscriteria opgesteld door de centrale overheid, laten we zien dat een extra monitor om luchtverontreiniging te meten de uitkomsten van vervuiling, gemeten door satellieten, verminderde met 2-3% en de handhaving van voorschriften voor luchtverontreiniging met 20% verhoogde. Om te verhelderen hoe deze vermindering van vervuiling tot stand kwam, construeren we een nieuwe dataset over handhavingsactiviteiten van stadsbesturen met betrekking tot luchtverontreiniging van 2010 tot 2017. We georefereren handhavingsactiviteiten en laten zien dat de toename de handhaving wordt gedreven door bedrijven die dicht bij die monitors zijn gevestigd, terwijl bedrijven ver weg met minder wijzigingen in handhaving te maken hebben. Deze effecten worden veroorzaakt door lokale ambtenaren die sterke prikkels hebben om de vervuiling te verminderen en die sterker zijn wanneer er beperkte mogelijkheden zijn voor datamanipulatie – wat suggereert dat informatie van betere kwaliteit de verantwoordingsplicht van uitvoerende ambtenaren kan versterken en de beleidsresultaten kan verbeteren.